Simulation and optimization methods to improve the management of resources and patients in health services. Application to emergency departments.

Marta Cildoz Esquiroz

Department of Statistics, Computer Science and Mathematics

Public University of Navarre

Advisor: Prof. Fermin Mallor Giménez

November 2019

Acknowledgements

I would like to acknowledge all the people who have helped me during this *journey*, especially to my family and friends.

First, I would like to express my sincere gratitude to my advisor, Prof. Fermin Mallor for his unevaluable continuous support, help, and guidance since I started in the research field developing my Master Thesis until the final stage of this dissertation.

I also wish to thank

- the *Department of Statistics, Computer Science and Mathematics* of the Public University of Navarre, in particular to Dr. Cristina Azcárate for her unconditional support.
- all my colleagues at my department at the Public University of Navarra who I shared many unforgettable hours with.
- Dr Pedro Mateo, from the University of Zaragoza, who initiated and accompanied me into the Java programming language world.
- Prof. Ana Póvoa, Dr. Inês Marques, Mariana, and all my colleagues at the Instituto Superior Técnico at the University of Lisbon for their infinity hospitality during my research stay that made the experience wonderful.
- Javier Gorricho, who works at the *Servicio de Planificación, Evaluación y Gestión del conocimiento del Gobierno de Navarra*, for providing us with the administrative data necessary to develop and validate our models.
- Isabel Rodrigo, who works at the *Servicio de Apoyo a la Gestión Clínica y la Subdirección de Procesos de Hospitalización y Urgentes, del Complejo Hospitalario de Navarra*, infinite source of real management problems arising in the Hospital Complex of Navarre.
- Javier Sesma and Iñaki Berrozpe for providing us with all the requirements that need to be incorporated to the physician schedules, and for validating and discussing with us the shortcomings of the algorithmic solutions.
- Amaia Ibarra specially, an experienced physician at the ED, who has been my travel companion in the analysis of the ED management problems. She has helped me to understand the work carried out by emergency staff and then to keep the mathematical models developed in this thesis connected with reality and therefore, making the results useful to improve the management of emergency services.

- the Government of Navarre for its PhD grant that allowed me to enter the exceptional world of research.
- the ORAHS community and their experienced professors and researchers for their stimulating atmosphere which encourages fellowship and intellectual communication.

Abstract

The aim of this thesis is to contribute to the sustainability of public health services by means of data analysis and through the development and application of Operational Research methods and techniques for modeling and analyzing real planning and management problems generally affecting the public health sector and Emergency Departments (EDs) in particular. The focus of the research is on the development of methods of analysis that will yield practicable solutions to improve the efficiency and quality of patient care and working conditions of the health staff.

A hospital ED provides medical and/or surgical care to patients arriving in need of immediate attention. The highly stochastic environment of these departments is especially difficult to manage due to the variability of the patient arrival rate, patient severity, and (material and human) health resource requirements. They also have to provide a 24/7 service, where physicians are required to work night, day and weekend shifts, and take on different assignments.

The research for this thesis covers two types of problem: the improvement of patient flow management and physician shift scheduling. Simulation techniques were selected to model the variable and stochastic environment of the ED. The resulting model includes seasonality in patient arrival patterns by level of severity, and mimics patient pathways through the ED, reflecting the resource consumption (including the medical staff) required for treatment. A guideline is provided for the construction of a mathematical model of the ED designed to overcome some of the shortcomings of oversimplified queuing theory models and capture some important issues that previous simulation models have overlooked.

The first part of the thesis addresses the problem of patient-to-physician allocation following triage. It offers a proposal for new allocation rules which prove to outperform the common cyclic allocation approach by taking into account a factor usually neglected by patient-flow management policies: i.e., the workload stress experienced by physicians, which is measured in real time using a method proposed and analyzed in this thesis. The stress score is used as the KPI to assess the performance of current patient-flow management policies and as a criterion for designing new ones. This thesis also illustrates the successful implementation of one of the proposed rules, from initial concept to practical application in the hospital. The tested allocation rule outperforms the current cyclic one, as demonstrated by using the simulation model and analysis of the real data gathered during the pilot test.

The second part of the thesis addresses the physician scheduling problem, which is a combinatorial optimization problem posing particular difficulty when all the constraints and

objectives observed in practice are considered. The problem is modeled by means of mathematical programming, and thus cannot be solved in practice by commercial software. This leads to the development of a new solution heuristic. A key feature of this algorithm is the greedy constructive phase, which is guided by solving a linear problem in combination with a memory structure. Initial good solutions are very quickly obtained, but they can be unfeasible in heavily constrained cases. The subsequent improvement phase combines a repair strategy based on variable neighborhood search with network optimization. This is the first proposal for such a strategy. A computational analysis and a real-case solution demonstrate the quality of the solutions and the good behavior of the methodology.

The research presented in this thesis fulfills the following objectives:

- To propose a quantitative framework (based on simulation models and their combination with optimization procedures) for the analysis of problems involved in the dimensioning and assessment of management policies in hospital emergency services.
- To develop a methodology for the real-time assessment of pending workload stress in physicians.
- To provide new patient-to-physician allocation methods with criteria including the workload and stress balancing across physicians, and patient service quality.
- To analyze alternatives to pure priority rules for managing the queue of patients awaiting initial emergency assessment by a physician or reevaluation following tests and/or diagnosis.
- To design efficient algorithms for solving the physician work-shift assignment problem taking into account all real ergonomic constraints while balancing the workload.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Importance and complexity of Emergency Departments (EDs)

Public health expenditure represents a very important part of national budgets in the developed world. In Spain, it amounted to 75,435.4 million Euros (€ 1,617 per capita and 6.24% of GDP) in 2018, having increased considerably since the end of the last century (in 2000, for example, it accounted for 4.85% of GDP with a total of 31,432.3 million Euros and 775 € per capita). The last decade, however, has seen a degree of stagnation and even some slight decline (for example, from 6.77% of GDP in 2009). In addition, public health services are facing a growing demand due to the aging and longer life expectancy of the population, as well as an increase in patient expectations and demands.

The Spanish National Health Survey of 2017 reveals the public's intensive use of health services: 85.8% of the population (91.4% of women and 82.4% of men) report having consulted a doctor in the last 12 months. Almost a third of the population (31.3%) was attended by an Emergency Service (ES) in the last year, the frequency being higher among children and the elderly overall (see Table 1.1) and higher for women than for men (see Figure 1.1). Emergency departments (EDs) are where most of the emergency health care (93.42%) is provided. The usage rate is significantly lower among the elderly, who have more need of home care and mobile units than the rest of the population (see Table 1.1). Some patients attend the ED several times a year, taking the average number of usages per patient to 1.82, in 2017.

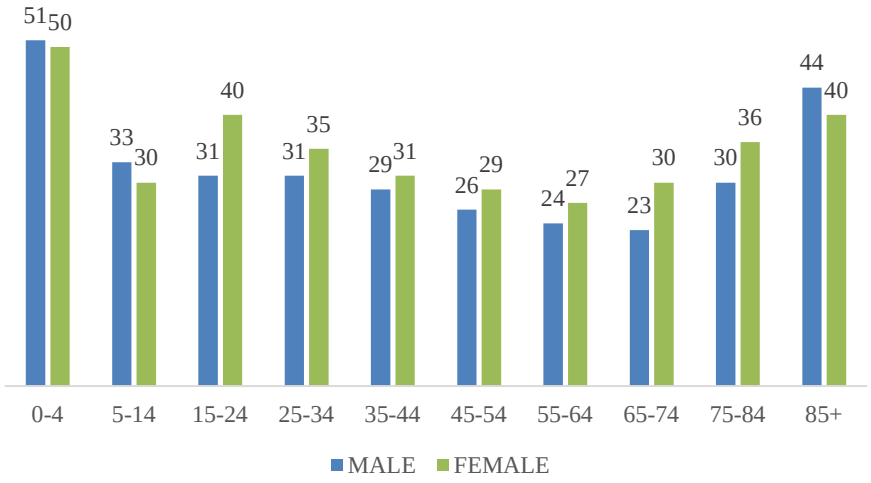


Figure 1.1 Percentage of the population that used ES in 2017, distributed by age. Data from the National Health Survey, 2017, prepared by the Spanish Ministry of Health, Consumption and Social Welfare.

The Spanish Society of Emergency Medicine estimated the total number of visits to hospital ES in 2018 at 28 million, the enormous economic impact of which can be estimated in terms of the average cost per visit for the Health Departments of some of Spain's Autonomous Communities. The cost estimates vary widely: while the order 731/2013 of September 6 of the Madrid Govt. Health Ministry estimates the cost of emergency hospital treatment without admission at 180 Euros (190 if it involves trauma), that of Galicia is estimated at 362 Euros. In any event, the total estimated cost of emergency visits exceeds 5,000 million Euros (more than 100 Euros per capita).

Table 1.1 Emergency Service usage rates, number of usages and location of service provision. Data from the National Health Survey, 2017, prepared by the Spanish Ministry of Health, Consumption and Social Welfare.

Age group	% of use of ES	Average number of usages	Standard deviation of the number of visits	% of Emerg. at home, workplace…	% of Emerg. Requiring a mobile medical unit	% of Emerg. at Emergency Unit
0-4	50.67	2.24	2.08	3.88	0.26	96.40
5-14	31.94	1.65	1.52	3.36	0.76	96.56
15-24	35.66	1.98	2.89	3.73	0.44	95.90
25-34	32.84	1.97	4.48	3.97	1.24	95.47
35-44	29.82	1.78	1.92	4.63	1.38	94.93
45-54	27.67	1.70	1.71	4.55	2.96	94.46
55-64	25.27	1.73	1.61	7.63	2.76	91.68
65-74	26.92	1.74	1.78	7.77	4.77	89.44
75-84	33.37	1.73	1.45	13.15	4.80	86.99
85+	41.37	1.66	1.24	19.94	9.26	78.29
TOTAL	31.27	1.82	2.39	5.92	2.26	93.42

The last 30 years have seen a growing trend ES usage, which has risen from 11% to 31% of the population (see Figure 1.2). This increase can be attributed to multiple causes (see [1]):

- The progressive aging of the population (particularly significant in Spain which has the second longest life expectancy and one of the lowest birth rates in the world), which leads to higher chronicity and dependency rates.
- The improvement in ES, both quantitative, due to the increase in the number of units, and qualitative, due to human and material resource improvements.
- Public demand for immediate solutions to health problems leads patients to use the ES to bypass waiting lists and gain direct access to a specialist consultant.
- ES provide the only available public health care assistance for almost two thirds of the calendar year (nights, weekends, holidays), which, in combination with the need for immediacy cited in the previous point, contributes to increase demand.

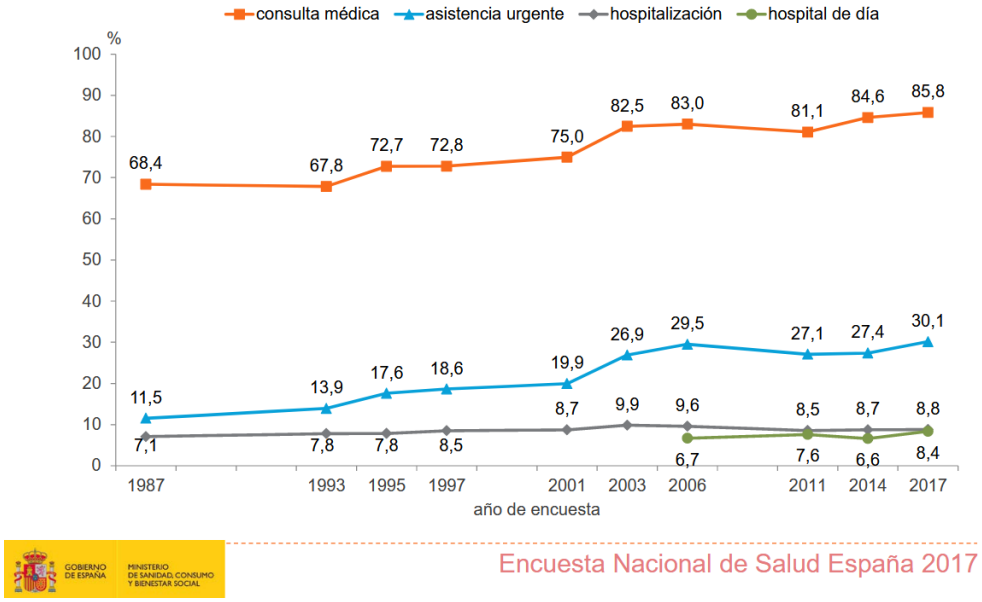


Figure 1.2 Use of health services 1987-2017 in the last 12 months by the Spanish population aged 15 years or over. Data from the National Health Survey, 2017, prepared by the Spanish Ministry of Health, Consumption and Social Welfare.

The control of ES demand is very difficult due to the unique nature of the sanitary emergency, which is defined by the World Health Organization as "*the accidental occurrence (unexpected or unpredicted), in any place or activity, of a health problem due to different causes and of varying severity, which generates the awareness of an imminent need for attention to the individual suffering the problem or his/her family*". This definition does not allow for an objective assessment of when to consider an unexpected health problem urgent and eligible for treatment in an ED, since it depends on the subjective assessment of the patient or someone in his/her presence, and it is their decision to request the services of emergency professionals. The American Medical Association also includes subjectivity in its definition of an emergency, defining it as "*the condition that, in the opinion of a prudent layperson (the patient, his/her*

family, or whoever assumes responsibility for the patient, requires immediate medical attention ". Both definitions are widely accepted by national health agencies. Therefore, the ES must admit every patient who considers that her/his health problem needs immediate attention.

ES in Spain and other countries with a free public health care system can therefore be considered as a public service that fits the premises of the tragedy of the commons [2]: free access, absence of restrictions on individual conduct, demand that exceeds supply and the inability of users to modify the rules.

The tragedy of the commons is a social dilemma, described, as we know it today, by Garrett Hardin [3] in 1968. As C. Navarro [2] points out in his thesis, "*the dilemma indicates that when common goods or resources are considered, with freedom of access and free of charge, each person, as a rational being, seeks to maximize her/his utility, so people are persuaded to use the aforementioned good unlimited, since the utility obtained from its use is always a positive figure, although the sum of similar behaviors by a large number of people will cause a deterioration in the aforementioned good, due to its excessive use, and therefore a long-term damage to society as a whole (and consequently to each of its members)*". Navarro investigates the introduction of co-payment as a way to induce more rational use of the ED and reduce its costs, thereby contributing to its sustainability.

The ES form a complex organizational structure, incorporating a wide range of professionals with multidisciplinary training, providing their services 24 hours a day, 7 days a week, and 365 days a year, holidays included. These professionals must be able to work under time pressure, since response time is crucial in the case of seriously ill patients whose prognosis depends on the nature of the clinical decisions and the time taken to reach them. The Spanish Society of Emergency and Emergency Medicine also highlights the essential role they play in time-dependent pathologies such as stroke, heart failure, heart attack or traffic accidents; and their decisive participation in catastrophes and multiple-victim incidents such as terrorist attacks. This activity, carried out in an unpredictable environment, cannot be planned or programmed, thus making the ES very difficult to manage.

The necessary coordination of the ES with other health care units further increases their complexity and hinders their effective management. The activity of these other units affects and is affected by the activity of the ES. The interaction of the ES with primary care services, intensive medicine services and other hospital departments is not always clear in organizational terms. The activity of the ES conditions hospital and ICU bed management and also surgical programming. Similarly, the activity of these other units determines activity in the ED, since it can lead to emergency patient blockage, and delay in hospitalization, ICU admission or performance of surgery, thus increasing the risk of congestion. Meanwhile, the referral of certain primary care patients to the ES further contributes to overcrowding.

1.2 Management problems in EDs and quantitative approaches of analysis

The demand increase described in Section 1.1 above has led (and will further lead, if nothing is done about it) to problems affecting the health both of ED patients, through longer waiting times and waiting lists, and of ED workers, who typically experience a higher rate of burn out than other healthcare professionals. This issue has great social impact and has attracted the attention both of the media and politicians. Problems of congestion and overcrowding in Spain's ES have also been raised by its national and regional ombudspersons and exposed in the book "*Hospital emergencies in the National Health System: patient's rights and guarantees*" published in 2015 [4]. Among its conclusions, the following stand out: "*Emergency management is the time management, so the dashboards should be able to establish the times and phases of patient care (traceability) while staying in the ED, as a measure to seek more effective care. The computer applications implemented in most hospitals do not take into account the specific requirements of urgent care*". The book calls for the use of modern information technologies and artificial intelligence to improve the management of the ES.

Informed and efficient resource management is therefore necessary to ensure the quality and sustainability of public health services. However, the usual planning and management methods are based on the experience of those in charge, most of whom have medical or related training, acquired in contexts that are no longer valid. As already stated, the health services have become extremely complex systems evolving in highly variable and unpredictable environments, and requiring management policy improvements which can only be achieved with the help of modern information technologies and decision-making science.

There are two categories of ED management problems; those arising from the relationship of the ES with the other health units with which it interacts to improve patient flow through the whole system (ICU, hospital, primary care); and those relating to resource and patient management in ES facilities.

The first category of problems affect the performance of the entire health system and must be addressed by means of interconnected information systems and coordinated decision-making, which, in the current context of Spanish public health, are far from being a reality. Internal ED management problems, on the other hand, can be tackled using ED patient records, and ED resource files. Such problems may be strategic, such as the layout of the facilities; tactical, such as human resource deployment and shift allocation; or operational, such as patient flow control.

This thesis proposes the development, application and implementation of quantitative methods to provide scientific support to the tactical and operational management of ES, using data analysis and decision-making based mainly on optimization and simulation techniques. This is

an Operations Research (OR) approach for solving real, complex problems, generally involving the allocation of scarce resources.

The use of OR methods to solve health problems is not new, as can be observed in the activity of the European working group ORAHS (Operational Research Applied to Health Services), specialized academic journals such as Health Care Management Science (created in 1998) and Operations Research for Health Care (created in 2012), and the health sections of the main OR journals (EJOR, Operations Management, Omega, etc.). Countries such as the United States, Canada and others in Western Europe (Holland, United Kingdom, France, Germany, Italy, Portugal...) have a long tradition both of academic research and collaboration between the university and health services for the analysis of real problems. In Spain, however, there are few such experiences and no stable collaboration of that nature.

The main OR tools that have been used to improve the management of ES in general and ED in particular, are Queuing Theory, Simulation, Mathematical Programming and Markov models.

Mathematical programming has been used mainly to address staffing and nurse- and physician-scheduling problems. When the mathematical models become too difficult to be solved to optimality using commercial solvers, then heuristics methods are developed. Among the most successful is the Tabu Search, which has been used successfully in several studies (e.g.[5]–[7]). Mathematical programming is also used in combination with forecasting models to study staffing and scheduling problems taking into account intraday variability in patient arrivals. A literature review of these problems is provided in Chapter 6 of this thesis.

Queuing theory has been extensively used to analyze ED patient flow. A rationale for its use is explained in Wang et al. [8]: "*Although analytical methods contain fewer details than simulation, and are based on simplified models, it could provide quick results and an opportunity to investigate system properties more efficiently under appropriate assumptions*". Many papers use Markovian models to study specific queuing problems in healthcare systems (a review of Queuing theory applications in healthcare is provided in Lakshmi and Iyer [9]). However, their appropriateness depends on the research assumptions and goals. When used to address complex ED management problems, queuing models generally show some deficiencies, mainly by usually assuming stationary arrival processes, simplifying the service process (when provided in several steps and, possibly, in different locations) and ignoring patient exits from and re-entries into the system. The discrepancy between queuing model dynamics and the real process dynamics has been analyzed recently in Azcarate et al. [10] for the Intensive Care Unit (which is arguably a simpler system for modeling purposes). Papers using queuing models to analyze the patient flow in EDs are reviewed in Chapter 5 of this thesis.

The simulation approach to the study of ED patient flow and staffing and scheduling problems overcomes the difficulties pointed out in relation to the queuing theory approach. In the words

of Kolker [11]: "*Process model simulation approach seems to be much more flexible and versatile. It is free from assumptions of the particular type of the arrival process (Poisson or not), as well as the service time (exponential or not). The system structure (flow map) could be of any complexity, and custom action logic can be built in to mimic practically any features of the real system behavior*". The ability to model processes in great detail motivated the development of a simulation model that captures all important influences on patient flow and ED management decisions. The resulting model is presented in the next Chapter of this thesis (Chapter 2). It is thanks to its capacity for "what if" analysis that simulation has become the choice of analytical tool in many contexts where the best of several alternatives must be selected. This approach is used in Chapter 4 of this thesis to improve patient-to-physician allocation after triage. However, it is its capacity for combination with other analytical tools, such as optimization, that makes simulation so powerful for the analysis of complex problems. Such a combination is used in Chapter 5 of this thesis to obtain optimal patient flow management policies. Reviews surveying the use of simulation in EDs include Brailsford[12], Günal and Pidd [13] and, most recently, Vanbrabant et al. [14].

Markov Decision Processes enable the modelling of ED patient flows and patient reneging. They are used in several papers for length-of-stay prediction and patient flow design. The review of problems and methods by Saghafian et al. [15] surveys these applications.

1.3 Introduction

Simulation techniques, as mentioned in the Introduction Chapter 1, are the best option for modeling and analysing EDs [14], since they can accommodate the desired level of detail and deal with the stochastic nature of ED queuing patterns. Thanks to these characteristics, simulation models are able to approximate real-life behavior, and yield reliable "what-if" analyses (e.g. [11], [16]). Simulation models also serve as a basis for system optimization using simulation-optimization techniques.

Computer simulation is therefore a widely-used tool in health-care studies, as shown in several comprehensive literature reviews, and, within the health-care domain, much attention has been given to the simulation modeling of EDs ([12]–[14], [17]). However, the generated output is valuable for the investigation of ED operations only if the model closely resembles the real system. Model design and construction is therefore of prime importance. The aim in this chapter is to describe all the steps required to build a credible and valid simulation model that can be used to analyze patient flow and resources. It will relate how a simulation model is developed for the Hospital Complex of Navarra (HCN) ED, and used, once validated, for the investigations proposed in the subsequent chapters.

Conceptual model design is a blueprint of the model that is to be built. While ideally independent of simulation software [18], it is nevertheless dependent on the selected simulation

method [19] which determines the modelers' particular perspective of the world . A conceptual model is a necessary prerequisite of model construction, as it aids understanding of the problem scenario, the definition of research objectives, and the identification of boundaries, inputs, outputs, construct components and their interactions. This is especially important when modeling a healthcare setting, particularly EDs, whose organizational complexity must be reduced to enable the fulfillment of the modeling objectives. One of the commonest approaches to this key task is through process flow diagrams.

The objectives, level of detail, and the generality of a model are interrelated. The greater the level of detail, the closer it reflects reality, but the less likely it is that the model will be generic. Generally speaking, ED modelling objectives focus on activities that affect hospital performance and therefore have a major impact on the level of detail and generality of the model. However, if the model is properly parameterized and sufficiently flexible, it can, with little modification, be reused to evaluate the impact of changes in the system (e.g., priority rules and demand changes) on ED performance as well as being transferable across different hospitals ('full model reuse').

Credible models require reasonable data. These provide the input; the more detailed the model, the more inputs it requires. ED modeling requires information (and data) from various sources, such as the hospital's information system (which collects data routinely in hospitals), interviews with staff, personal observations on site, etc. Considered and careful analysis of these data is an important phase in the development of most simulation models.

Various simulation methods can be used for building hospital ED models. The most commonly used are Discrete Event Simulation (DES), System Dynamics (SD), and Agent Based Simulation (ABS). This chapter describes the construction of a DES model. These have a long history and are commonly used to model systems that change states dynamically, stochastically, and at discrete intervals. This methodology is particularly powerful in systems with a strong queuing structure, such as the ED, where patients (entities) compete for resources, since it consists of tracking entities that change their state within a system. DES software also offers great flexibility, since it supports different detail requirements, enables easy modeling of stochasticity (random emergency arrivals, length of consultation, etc.), easy programming of complex queuing mechanisms, and a visual representation of patient flow. All discrete event hospital simulation models include entities and attributes (generally patients), resources (physicians, ancillaries, nurses, X-ray machines, etc.) which can change their state, a network of processes representing the interaction between entities and resources, and input and output variables.

This chapter starts with a description of the conceptual modeling of an ED including its main elements; the relationships among them are described in Section 2.1. Section 2.2 details the necessary model input data, and the data collection and estimation process. Sections 2.3 and

2.4 explain the implementation phase and the model verification and validation, respectively. Finally, Section 2.5 details the construction of the simulation model for the HCN.

1.4 Objectives

The overall aim of this thesis is to develop methods and algorithms for operational and tactical decision making aimed at improving the running of hospital ES, from both a patient and personnel perspective.

This overall aim is broken down into the following specific objectives:

A. To propose a quantitative framework (based on simulation models and their combination with optimization procedures) for the analysis of problems involved in the dimensioning and assessment of management policies in hospital ES.
B. To develop a methodology for the real-time assessment of pending-workload stress in physicians.
C. To provide new patient-to-physician allocation methods with criteria including the workload and stress balancing across physicians, and patient service quality.
D. To analyze alternatives to pure priority rules for managing the queue of patients awaiting-initial emergency assessment by a physician or reevaluation following tests and/or diagnosis.
E. To design efficient algorithms for solving the physician workshift assignment problem, taking into account all real ergonomic constraints while balancing the workload.

The research carried out in this thesis aims to provide solutions to classic problems involved in the allocation, planning and management of ED resources in order to mitigate some of the adverse effects of congestion and overcrowding on both patients and personnel. The implementation of the results obtained in this thesis should be beneficial for:

- Decreasing waiting times and the proportion of patients not seen within the target time to first contact.
- Decreasing the risk of poor clinical outcomes due to delays in diagnosis and initiation of treatment.
- Reducing the number of patients leaving the service without being seen.
- Balancing the physician workload, thereby reducing burn-out and absenteeism.
- The prevention of attacks on health personnel, some (unreasonably) due to long waiting times or non-application of the FIFO discipline.
- Continuity of care
- Reducing the service costs incurred by inefficiency of care.
- Improving overall productivity by reducing the time spent in the ED by patients and those accompanying them.

In summary, the thesis, based on a multidisciplinary research framework (quphs group, www.unavarra.es/quphs) for the analysis of real problems, aims to contribute to academic advances in solution of these problems, while making a real impact in the improvement of ES.

1.5 Structure of the thesis and main results

The thesis is structured as two initial chapters and another six divided into two parts, each dealing with a different problem. It ends with a chapter on conclusions and possibilities for future research (Chapter 9) and a list of references. Eleven annexes are also included.

Chapter 1 presents the research setting: hospital ES, with a focus on their economic importance and benefit to the public and particular emphasis on their complexity and management difficulty. This chapter also describes the general and specific aims of the thesis showing how they relate to real problems.

Chapter 2 describes all the steps required for the construction of an ED simulation model. The model incorporates all the structural (conceptual) and stochastic elements required to represent a valid ED, for use in the analysis of different patient flow management policies and changes in resource availability. The model, validated by physicians of the Hospital Complex of Navarra (HCN) ED, is used to perform various investigations in subsequent chapters.

Part I of the thesis occupies chapters 3, 4 and 5, which deal with the research carried out to improve the flow of patients through the ED. The approach encompasses two perspectives: that of the patient, by trying to minimize waiting times and access time to first contact, and that of the physician, by trying to achieve an equitable distribution of the workload in order to reduce stress and burnout.

A review of the mathematical and medical literature revealed no method for the real-time measurement of pending-workload stress in physicians. A specific method therefore had to be developed. This involved the definition of workload stress factors, the selection of workload scenarios for stress assessment, the design of surveys to elicit the opinion of ED physicians and the proposal of a mathematical stress-assessment model with parameters estimated by the statistical analysis of the physicians' answers to the surveys. The presentation and scientific rationale for the method, and its real-case application are presented in Chapter 3.

The control of patient flow through the ED has two stages, the allocation of patients to physicians following priority-labeling during triage and the management of patients awaiting initial assessment or possible reevaluation following tests and/or diagnosis.

Patient-to-physician allocation is analyzed in Chapter 4, where proposals are made for different allocation rules aimed at different objectives, including the minimization of physicians' stress, equitable distribution of the workload and the minimization of access time to first contact. The

chapter also describes the practical implementation of one of the rules, from its design and evaluation using the simulation model, through to its validation by the analysis of real data gathered during the hospital trial. The process also requires convincing managers, computer implementation and staff training.

Chapter 5 investigates the management of patients awaiting first or subsequent treatment by a physician. Accumulative Priority Queuing (APQ) models have been analyzed for the management of such queues which include priorities, time constraints and reentry into the service, but only from the queue theory perspective for M / G / k models. The results of the research carried out help to identify the circumstances in which APQ models outperform pure priority rules based on patient severity and type of consultation, and reach the optimal APQ policy by means of simulation-based optimization. It should also be noted that the proposal and analysis of a variant of APQ shows it to be very similar to that used in ED management in practice.

The second part of the thesis, presented in Chapters 6, 7 and 8, is dedicated to solving the problem of ED physician shift assignment. This is a combinatorial optimization problem entailing particular difficulty due to its size and the large number of constraints and objectives to be considered. Particularly relevant is the modeling of preferences and equitable workload distribution. The planning horizon considered in most studies found in the literature is usually short, typically varying between two and four weeks. However, the Spanish workers' statute (BOE-A-2015-11430) stipulates that workers must be informed of their entire shift calendar for the following year before the end of the current one. Thus, given that the motivation of this research is to solve the shift-planning problem in a practical Spanish context, the problem we face is much larger than those discussed in the literature. Chapter 6 presents the mathematical model of all the constraints and objectives that arise in practice and formulates an integer linear programming problem. This includes four types of restrictions 1) demand and capacity restrictions 2) workload restrictions 3) equitable distribution restrictions and 4) ergonomic restrictions. Chapter 7 presents a new heuristic algorithm developed to provide very good solutions to this problem within a reasonable timescale. The algorithm consists of two phases, a randomized greedy construct and a subsequent improvement which combines a Variable Neighborhood Search and Network Flow Optimization. Chapter 8 presents an extensive computational analysis of the algorithm in real-based scenarios. It is shown that commercial software fails to provide an optimal solution to problems of similar size to the real one in reasonable times (not even in a week). It also includes the results of a real-case application.

Since each part deals with a different problem relating to the management of ED resources and patients, a separate literature review is included for each. In the case of part I, which deals with the analysis of patient flow, separate reviews are made of the problems treated in each of the three chapters. As a result, this thesis does not include a specific global literature review section or chapter.

The thesis includes 11 Appendices. The first five relate to Chapter 3 of the thesis, which proposes a method for the real-time assessment of physicians' stress. They contain the instruction sheets for completing the questionnaires to elicit expert opinions in the real-case description (Appendix A and Appendix D), example questionnaires (Appendix B and Appendix E), and mathematical details for calculating the expert consistency index (Appendix C).

Appendices F, G, H relate to patient flow management problems (Part I): Appendix F presents the notation and acronyms used in this part of the thesis, Appendix G presents the survey used to assess the outcome of the pilot testing of the new allocation rule in the HCN previously evaluated by means of simulation (Chapter 4); and Appendix H provides additional numerical results from Chapter 5, which studies patient management under the APQ discipline.

The three last appendices relate to Part II of the thesis: the notation and acronyms for the scheduling problem are presented in Appendix I, the mathematical modeling of all constraints and objectives of the problem is summarized in Appendix J, while a Linear Programming model for solving a general covering problem as a basis for the proposed heuristic is summarized in Appendix K.

Chapter 2 ED Simulation model

2.1 Conceptual modeling

The development of the simulation model should be oriented towards the research objectives for which it is intended. Thus, most studies of health systems, including EDs, examine the performance of the system in terms of waiting times, level of occupation, usability of resources, patient satisfaction, etc.

2.1.1 Main elements

The first step is to identify the main elements of the system under analysis. The principal elements influencing global ED performance are:

- The patients. These are the system entities entering and leaving the ED. Analysis of patient arrival patterns is very important when using the simulation.
- Patient flow through the ED service. The routing of patients within the ED considers all the processes they undergo during their sojourn and depicts the relationships between the different elements considered. It includes possible decision-making points throughout the entire care process.
- Resources.
  - Personal resources: physicians, nurses, administrative staff, specialist physicians, orderlies, etc. Nurses may be employed in the triage process or as general nurses in the patient care circuit. Physicians fall into different categories based on their ability to work without supervision, and this ranking may influence resource allocation and process time. Some injuries in ED patients require treatment by a specialist physician who may be permanently employed within the system or across several different hospital departments.
  - Movable resources (not including personnel such as orderlies): wheelchairs and stretchers – if these are limited-, elevators, etc. These resources, as well as orderlies, are very important components of the healthcare system; particularly in EDs where urgent cases, unable to move independently, may arrive unaccompanied and therefore require stretchers or wheelchairs.

- Layout and principal facilities. These include rooms where physicians receive patients, resources such as X-ray machines, scanners, examination boxes, resuscitation rooms, etc.

2.1.2 Patient arrival patterns

The analysis of arrival patterns is fundamental for the simulation model to reliably reflect reality. It is also very important to establish distinct categories of ED patients, since they are treated in different areas of the ED, have different care needs, and are evaluated differently (e.g. different target times to contact with a doctor).Many studies have documented the heterogeneity of arrival rates across different patient categories (see for example [20]–[22]). The categorization of patients can include not only their priority but also their mobility, age, or any other characteristics influencing the processes and resources required for their treatment.

The medical literature shows that ED patients arrive randomly, and are, by their own definition, not eligible for scheduled care. Their arrival patterns are modeled by Poisson processes, which are characterized by lapses of time between consecutive, exponentially distributed random events. These processes are not homogeneous, however, since the rate of arrivals is not constant over the course of the day, as demonstrated in several studies ([20]–[23]). These time-dependent arrival rates, which can be regarded as non-homogeneous Poisson processes (NHPP), have been well-studied (see e.g., [24]–[26]). Thus, arrival patterns vary throughout the day according to a Poisson probability distribution, as expressed by the following equation:

$$P(X = k) = \begin{cases} \dfrac{e^{-(\lambda t)}(\lambda t)^k}{k!} & , si\ k = 0, 1, \dots, n, \dots \\ 0\ other\ case \end{cases}$$

where k is the number of arrivals (event occurrences), e is the basis of the Neperian logarithm, and λ is the average number of events per unit of time. In this distribution, the mean and variance coincide with this last parameter.

The quantitative analysis in this step consists, first, in the application of statistical methods to determine the influence of different factors in the arrival rate, then, in the calculation of the mean arrival rates per hour across the study period, generally the whole year, based on historical data and accounting for previously-detected time effects. As well as intraday seasonality (rates fluctuating throughout the day with peak periods are widely accepted), there can be intraweek or yearly seasonality [27]–[29]; that is, differences in arrival rates according to the type of day (e.g. holidays, working days, etc.).Furthermore, seasonal effects in arrival rates are inversely proportional to patient severity[29], and can even be statistically non-significant for the most severe patients. Therefore, less severe patients – whose injuries are not strictly urgent- are likely to access the service when their family, social and work circumstances permit[30]. Thus, it is necessary for the simulation model to calculate the arrival rate, λ, of a

patient of category i at a time of day, t, based on the affecting parameters. For example $\lambda_{i,j,k}(t)$ where j is the type of day (day after a holiday, holiday, etc.) and k is the period of the year (by season or injury seasonality).

2.1.3 Flow diagram

The exact classification of all ED patient flow processes is impossible, since access to the system involves a complex series of decisions, tasks, and interactions with ED and hospital staff. They also vary from patient to patient based on severity and diagnosis. However, a general flow process for the "typical" ED patient can be determined as shown in Figure 2.1 and the following processes are typical for an ED patient:

- Critical patients arriving by ambulance or helicopter are rushed to the resuscitation rooms and treated immediately.
- A walk-in patient stands in line for mini registration at the "admission desk", where basic details, such as name, birth date, and social security number are collected.

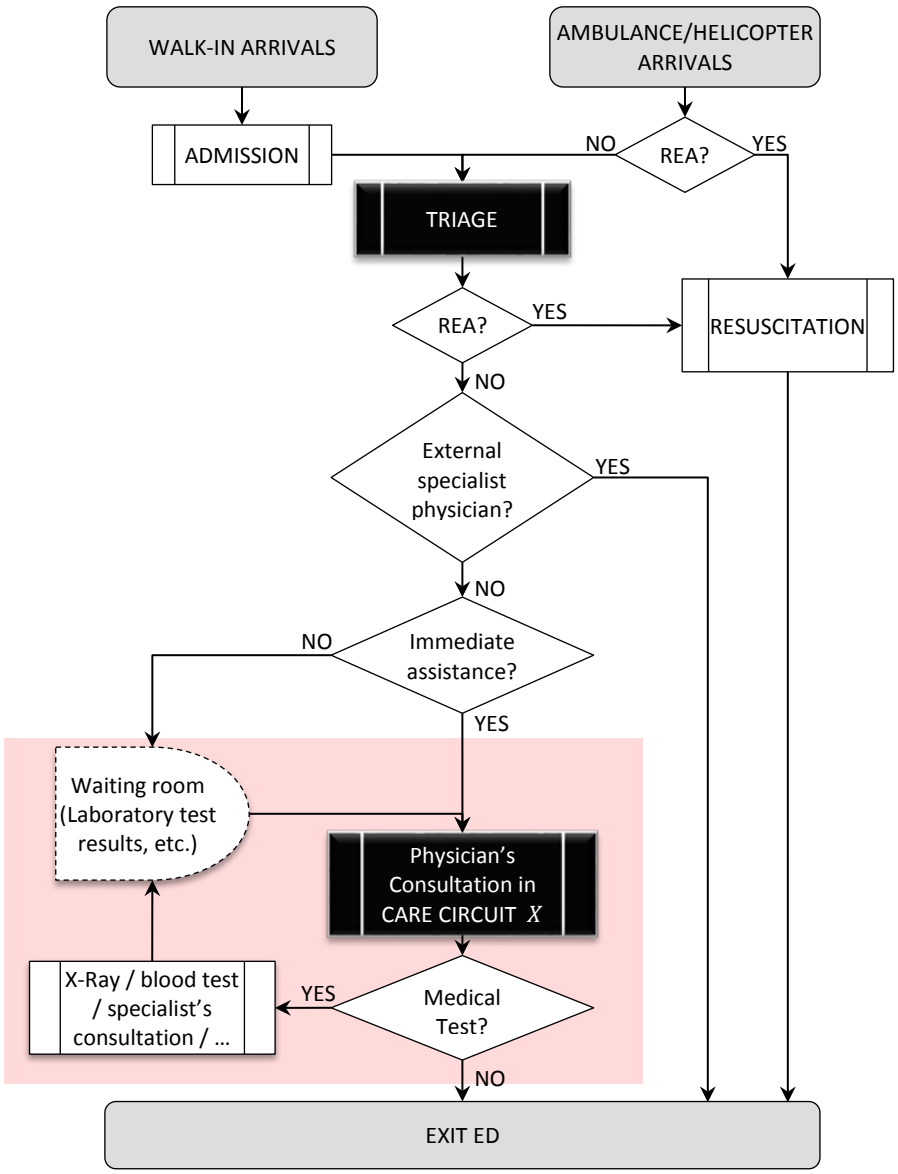


Figure 2.1 General flow process for a "typical" ED patient.

- Non-critical patients arriving by ambulance or helicopter and walk-in patients (after registration) undergo a triage process. Triage nurses call the patients in order of arrival, assess them and classify them by category (level of urgency) based on their symptoms, in order to ascertain their treatment priority. There are several triage protocols, which may include performance goals, such as the percentage of patients who must be seen by a doctor within a target interval. This will be explained in more detail in Part I of this thesis. Sometimes an ECG scan is performed during triage (as reported in several papers, e.g. [31])
- After triage, some patients need immediate treatment and are assigned to a room under the care of a nurse until the doctor arrives. Patients not needing immediate treatment

remain in the waiting room, which can be common or specific for each type of patient, until a physician or room is available for treatment to start.

- During treatment, the patient is evaluated by a physician, who may order clinical tests, such as blood tests, X-rays, scans, or a specialist consultation and reevaluate the patient and test results before allowing discharge.
- Patients with reduced mobility must be transported by logistics personnel.

EDs must operate at a very high level of efficiency to handle their typically large daily patient loads. Thus, at some points of the flow, decisions must be made to determine the order in which patients are treated; which physician should be assigned to each patient, etc. All these issues will be addressed in detail in Part I of the thesis.

2.1.4 Staffing and resource level

The principal resources considered in ED systems are: examination and procedure (e.g. X-Ray) rooms and staff such as nurses, logistic personnel, administrative assistants, and physicians, each of which may have associated costs as well as different number scheduled for each period of the day (defined shifts). It is important to specify how many staff are scheduled for each zone and its respective procedures. That is, the number of triage nurses scheduled, the number of nurses in each care circuit, etc.

Physicians may be senior physicians, medical residents in their first year of training or other medical residents in training. They differ in terms of whether they can work without supervision, or need to be coupled with a senior physician, etc. and with respect to the time they need to complete the consultation with the patient.

External staff, such as external doctors, referred to above as “specialist physicians”, are not usually considered in ED studies. ED physicians may call upon a doctor from an in-patient unit for clinical examinations or consultation, or on a trauma specialist for specific X-Ray imaging tasks.

2.1.5 Layout

It is important to consider the layout of the ED, which can affect processes, resource capacity (number of examination rooms, etc.), movements within the system, and patient flow, and can be organized in different ways. For example, there are usually physically separate triage rooms for different arrival modes (ambulance or walk-in) which affect the process time and transportation of each patient. Another possibility is that patient care may be organized into different care circuits; a process known as “streaming”. These streams don’t often share resources (physicians, nurses, ancillaries, rooms, etc.), as will be explained in more detail in Part I of this thesis. Finally, there are also two popular ways to organize the patient’s care process depending on the layout and the ED:

- A room is assigned to a patient who stays in that room until treatment is complete and the doctor orders discharge. The doctor sees the patient for an initial assessment and again once tests and a complementary diagnosis have been carried out and once the results are ready (if tests have been ordered) for reevaluation prior to discharge (processes shaded in red in Figure 2.1). Under this layout and treatment structure, the number of examination rooms limits the number of patients that can be treated simultaneously.
- The patient does not stay in the room during the whole treatment but only when called by the physician for initial assessment and possible reevaluation following tests and/or diagnosis. The patient spends the rest of the time in the waiting room or undergoing clinical tests.

2.1.6 Service time distribution

Some patients may require a number of different resources (e.g. physician, nurse, and examination room) which need to be coordinated and simultaneously available before the process can begin. This information is required as well as the estimated service times.

In the event of some operational process data being unrecorded, incomplete or unreliable for time distribution purposes, several estimation procedures are possible. Simulation-optimization approaches are sometimes proposed to obtain a good set of input parameter estimates for the simulation model (e.g. probability distributions of service times). Kuo et al. [22], for example, used the available data for the time lapses between the start of two different services for each patient, and, assuming a Weibull distribution, which can fit many continuous functions on the positive real line, they estimated process times using a simulation-optimization approach. Service times are frequently modeled based on the judgment of a team of ED experts, who often use a triangular distribution model to estimate the maximum and minimum values ([32]–[34]). The most likely value was set at $(minimum + 1/3(maximum - minimum))$ to reflect the positive skewness that is generally present in processing times [34]. Otherwise, if possible, patient processing time in specific areas of the ED is gathered by sampling techniques [21] and an appropriate random distribution can be estimated for each set of data obtained.

2.1.7 Patient pathways

The probability of requiring a given step of treatment varies across patient types and degrees of urgency. It is therefore fundamental to model the defined sequence of each patient type within the ED system based on emergency care phases. By analyzing their historical records, patients can be grouped, and assigned to a route once their routing probabilities have been calculated.

Typical ED routes – not including the resuscitation room – are: for patient is sent directly to a specialist physician within the hospital, patient is discharged after initial consultation with the

physician, or patient is sent initial consultation with a physician for treatment/medical tests before being reevaluated. The medical tests can include: X-Ray, blood test, urine analysis, CT scan, assessment by a specialist, etc.

2.2 Data collection

Data acquisition is a crucial part of simulation modeling and involves various sources. The representativity of the input data as well as the statistical accuracy of the model influence the value and usefulness of a simulation model as a tool for system research and analysis of possible ED management alternatives. It is important to use reasonable data to build credible models.

The necessary input parameters/data previously described are arrival rates, probability distributions of service times, available resources, doctors' and nurses' schedules, etc. The input data for any model depends on the required level of detail. Modern information technology helps in this respect by routinely collecting an extensive amount of data. An ED's computerized patient tracking system usually records the time each patient spends in the various emergency care phases: arrival time, beginning of triage, beginning of initial assessment by the physician, medical test requests, etc. Although most of the data is already being collected and stored in the ED's computerized databases, some, such as processing times, may be missing or unavailable, and thus compromise the required level of detail. As already stated, when sufficient empirical data is not available, information sometimes has to be obtained by on-site sampling or from estimates by so-called subject matter experts (SMEs) [35], who can typically provide estimates of the times required for the different processes included in the model.

Therefore, this data collection phase also combines data from observations and interviews with experts and practitioners, who can provide holistic insights for various system issues. Healthcare systems contain a high level of complex social interaction particularly at decision points. Interviews and observation contribute greatly to a better understanding and an accurate modeling of work flow in the healthcare facility. They can also fill information gaps that cannot be addressed solely with numerical data or the hospital's information system. Interviews with healthcare facility senior managers, moreover, are essential for ensuring correct justifiable decisions and examining potentially effective solutions for the future.

Finally, it should be stressed that, for the sake of reliability, data cleansing procedures are required prior to the extraction and analysis of any data set from patient records to estimate patient arrival patterns, patient groups, etc. This is all the more essential given that ED data are often taken in critical situations, under the pressure and stress caused by the patient's condition upon arrival. Systematic errors, duplicate records, incompatibilities (negative processing time values, etc.) have a great impact on data quality.

2.3 Implementation

The implementation of the simulation model is often referred to as model translation, as it combines the conceptual model validated by system facility experts and the results of analysing patients' historical records and data sampling to obtain a more detailed, complex and executable simulation model, which can be used to investigate the impact of possible decisions and alternatives (i.e., "what if" scenarios) to foresee the consequences of the decisions adopted by the decision makers.

The simulation model can either use code programming, or a simulation software package, which provides the tools that are typical and essential for certain modeling, as well as graphical visualization tools, which can be used as a communication platform in order to validate the model with the experts working within the real system and increase its credibility .

Modular implementation is useful for verifying the computerized model ([36], [37]) by executing each part separately to check output consistency against actual behavior in each case. This, together with a proper parametrization of the model processes, facilitates model design adjustments to test their impact on the system and adequacy to different hospitals.

2.4 Verification and validation

Verification ensures the correct transformation of the conceptual model by accurate computer programming and implementation, i.e. debugging the computerized model, while the validation process ensures that the simulation model is an accurate representation of the system under investigation [38] and suits its purpose [39].

The verification process begins with the generation of simulated arrivals to the system for comparison with historical arrival records. This is followed by statistical tests for equal arrival rates and equality of variance. Then, to verify the model logic, ensure that patients follow the correct care pathway as expected, and generate the desired spread of patient types and priority levels across the workload, the historical and simulated data relating to the percentage of patients following the different routes within the system must be compared.

The various simulation model validation techniques used throughout the design and development of the simulation model include those described below [35], [40]–[42]:

- Face validity (mentioned in [37]) is obtained through collaboration with the system's experts (ED workers, managers, specialists, etc.) who evaluate the model concept and compare its output with real-world system behavior. This involves their evaluating and interpreting the results to determine whether the usage rate, average number of due

patients, occupational level, etc., reflect their own experience in the field. This ensures that the simulation model adequately depicts reality ([43], [44])

- Animation (suggested in [45]) enables graphical visualization of the movement of patients, physicians and other medical staff through the ED in order to check whether patient dispatching rules and individual routings are as expected, as well as enhancing the credibility of the model. It helps experts such as ED managers and senior physicians to validate the accuracy of the model's logic and to check intermediate output values such as queue lengths and waiting times between processes.
- Historical data validation uses the system's empirical data to build the simulation model and derive several metrics, such as length of stay across patient types, access time to first contact, etc., for comparison with the simulation-derived measures. The comparison can be performed by graphical ([46], [47]) or statistical means such as the Chi-Square test to check the compatibility of time distributions [48].

Finally, this step also involves determining the simulation run length required for accurate estimation of the desired measure.

2.5 Construction of the Hospital Complex of Navarre (HCN) simulation model

The ED of the HCN, which is located in Pamplona, is staffed 24 hours per day, assists a population of half a million and has more than 140.000 annual users. This section describes the construction of the simulation model for this ED, beginning with the information sources, continuing with the conceptual model design and statistical analysis, and ending with the implementation and validation of the model.

2.5.1 Information sources

The hospital data base

The hospital administration database holds the electronic records of all patients who used the ED over the period 2014-2016. They total over 420.000 patients, each associated with more than 39.000 possible data fields and more than 550.000 complementary diagnostic test requests, each associated with over 700 data fields. Also included are the arrival times, number of physician consultations, medical test requests, and illness and acuity descriptions, among others (see Figure 2.2).

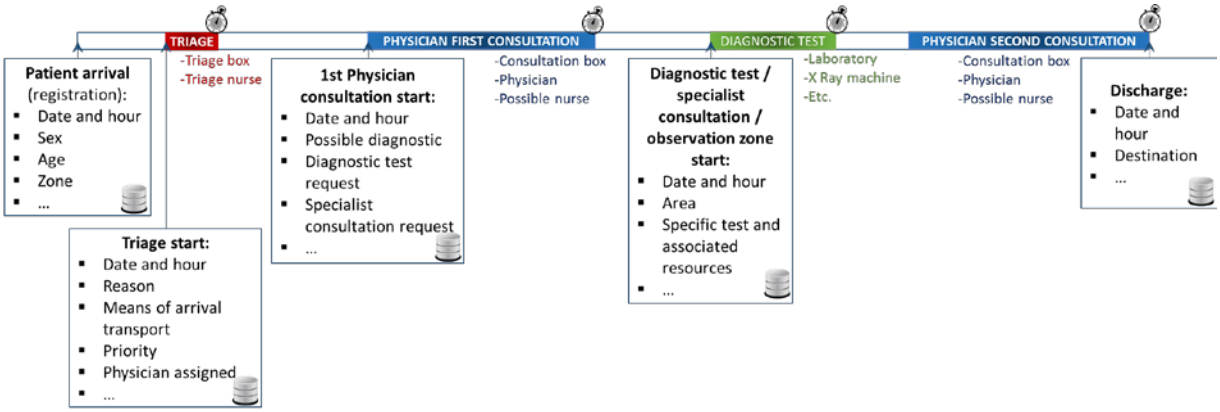


Figure 2.2 Hypothetical care process automatically recorded in the HCN database.

These data facilitate patient clustering (by sex, age, reason, diagnostic, priority, etc.), the estimation of arrival patterns across patient types, and patient pathways through the ED (including probabilities of branching at different points of the flow including the probability of discharge after the first consultation, specialist consultation requests, and medical test requests).

Interviews with staff and personal observation

ED staff interviews and observation during the workshift enable a better understanding and more accurate modeling of patient flow in the healthcare facility. They also enable modelers to see how physicians manage their workloads and allocated patient portfolios, since there are sometimes gaps in their self-reporting due to their having interiorized tasks to the point of performing them automatically.

Recording of processing time “in situ”

The duration of the patient’s different consultations with physicians, triage processes, medical tests processes, etc. are not recorded in the hospital database and therefore had to be recorded in situ and complemented by directly asking physicians and other staff members. Times for the different processes were recorded throughout each day for a whole month.

2.5.2 Conceptual model design and data analysis

Main elements

The main elements considered in the simulation model and used throughout this thesis are patients, patient flow through the system, human resources such as physicians, administrative staff, and triage nurses and exploration rooms.

Flow diagram

Patient flow in the HCN is similar to that shown in Subsection 2.1.3. The triage system is used to stratify patients into five different levels from most critical (1) to least critical (5). As in many other EDs, patient care is organized into two different circuits: one for the most critical, i.e., circuit B (CB), and another for the less critical, i.e., circuit A (CA) (see Figure 2.3). Thus, once classified by degree of urgency in triage, patients are assigned to one of the two care circuits for treatment: priority 1 (P1) and priority 2 patients (P2) are assigned to CB, priority 4 (P4), and 5 (P5) are assigned to CA, and priority 3 patients (P3) may be treated in both circuits. Then, each patient is assigned within the respective circuit to a specific physician, under whose authority he/she will remain throughout his/her stay in the ED system. The patient then waits in the waiting room of his/her respective care circuit until called by the physician for the initial consultation. After this, the physician may order medical tests or discharge the patient to one of several destinations such as an in-patient ward; his/her own home, or another hospital. Unless discharged to one of these destinations, the patient undergoes the requested medical tests and waits in the waiting room until the results are ready, and reassessment by the physician can take place prior to discharge. Patient flow management will be explained in more detail and addressed in Part I of this thesis.

Each circuit has dedicated resources such as physicians, nurses, waiting rooms, exploration rooms, etc. and both operate under the same management policies. Thus, after triage, they operate as two independent EDs. Figure 2.3 represents the general flow process for a patient in the HCN ED; delays, labeled “waiting rooms”, may be due to medical test results not being ready (laboratory blood tests), no physician being available, etc.

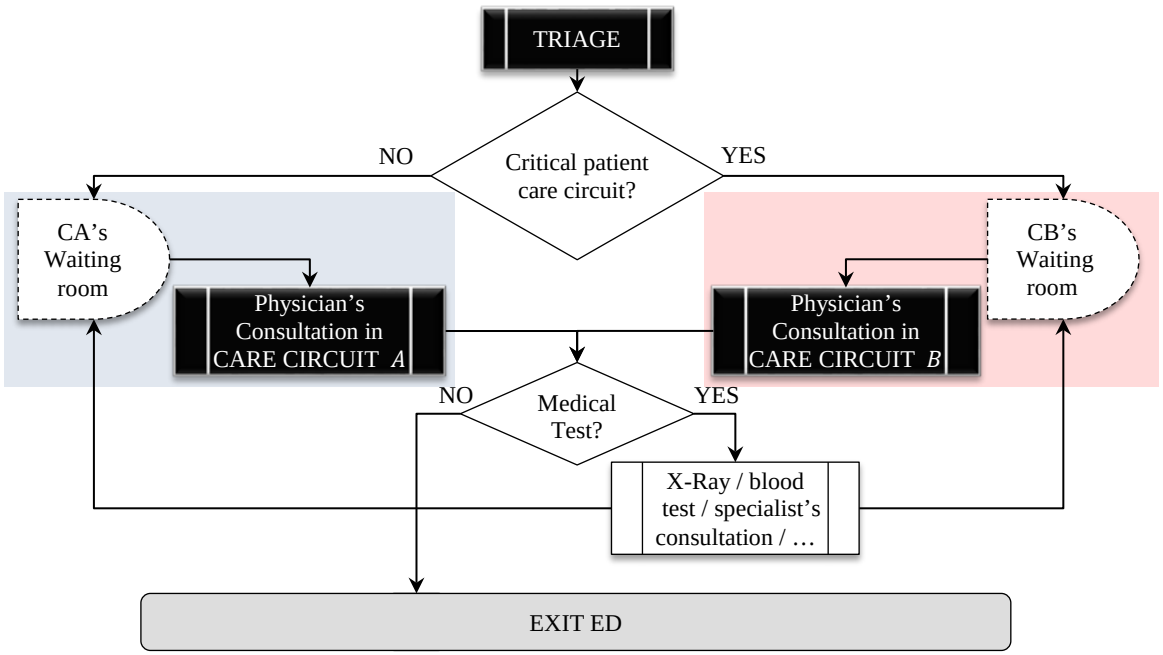


Figure 2.3 General flow process for a “typical” ED patient in the HCN

Patient arrival patterns

This part of the study began with a statistical analysis of arrival homogeneity over time and the influence of different factors, such as day of the year, day of the month, day of the week and type of day, patient priorities, and patient mobility. Using the electronic records of all patients presenting to the ED over the period 2014-2016 and the perceptions of ED workers in the HCN, the types of day were classified into holidays, days after a holiday, days before a holiday and working days not included in previous categories, and days in the week of San Fermin. Finally, the statistically significant factors for arrival rates were found to be patient priority, hour of the day, type of day (holiday, day after a holiday, and working days not after a holiday), and month of the year. Patient arrivals were modeled as a non-homogeneous Poisson process (NHPP) for each type of patient [49], with the intensity of arrivals $\lambda_{i,j,k}(t)$ depending on patient priority i, month of the year j, type of day k, and hour of the day t. This seasonality, also observed in other studies (e.g. [50]), depends inversely on the patient acuity level, such that lower acuity is related to higher intraday and intraweek seasonality [29], as stated in previous Subsection 2.1.2. In this case, there is no seasonality even for high priority patients across years or types of day. As an example, Figure 2.4 represents the hourly arrival rate of less critical patients (priority 4 and priority 5) throughout the day (0:00-23:00).

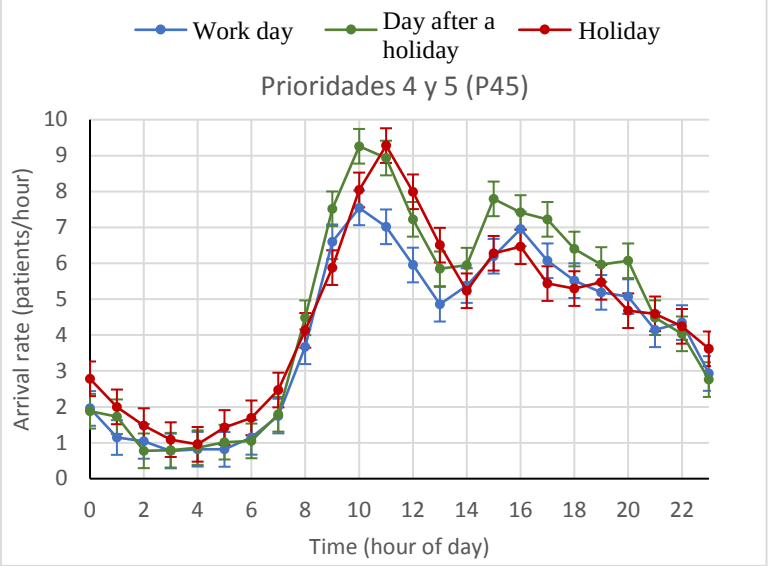


Figure 2.4 Average hourly arrival rate for less severe patients (priorities 4 and 5) across types of day: normal work day, holiday, and day after a holiday.

Staffing and resource level

This stage of the study is to determine the number of scheduled resources of each type mentioned in the previous Subsection “Main elements”. The main resources involved in all the issues addressed in this thesis are physicians, whose working conditions are very important for good quality of care, and examination rooms. Therefore, this study not only indicates the

number of physicians (each with an associated examination room) scheduled for each workshift and type of day, it is also flexible enough to enable the insertion of any number of physicians with a view to determining how many need to be scheduled for each workshift, (see Figure 2.5). The numbers of physicians scheduled are summarized in Table 2.1.

Table 2.1. Current number of physicians per shift, per day in the HCN.

	Working day	Day after a holiday	Holiday
8:00 – 15:00	14	16	10
15:00 – 22:00	11	13	10
22:00 – 8:00	5	5	5

The type of physician (first-year resident, general resident, or board-certified physician) is taken into consideration. The first type cannot discharge a patient without the supervision and approval of a board-certified physician; general residents can assist patients without supervision but may, in some cases, require a board-certified physician, depending on the priority of the patient, and the last type can assist any type of patient without supervision.

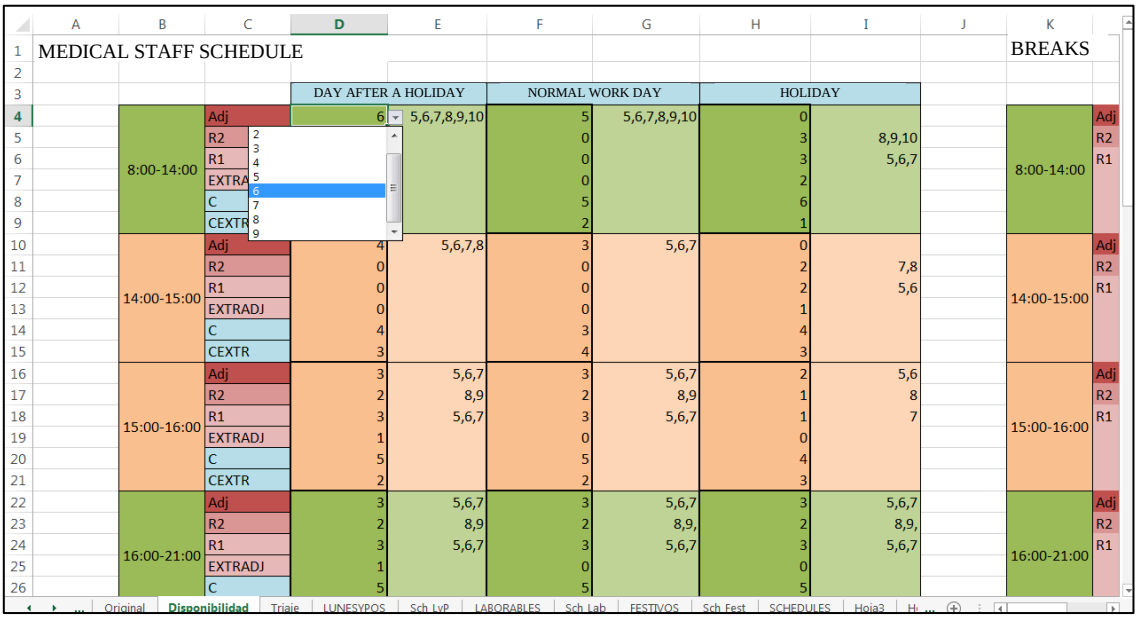


Figure 2.5 Work shift schedule and staffing template for each type of day in the simulation model.

Layout

ED layout is considered in order to check for resource limitations in terms of examination rooms (5 or 6 depending on the day for CA and 9 for CB), or triage boxes, and for the design of movement studies, etc. It was also used to design a 3D animation imitating the real system. Figure 2.6 shows a representation of the ground floor of the ED.

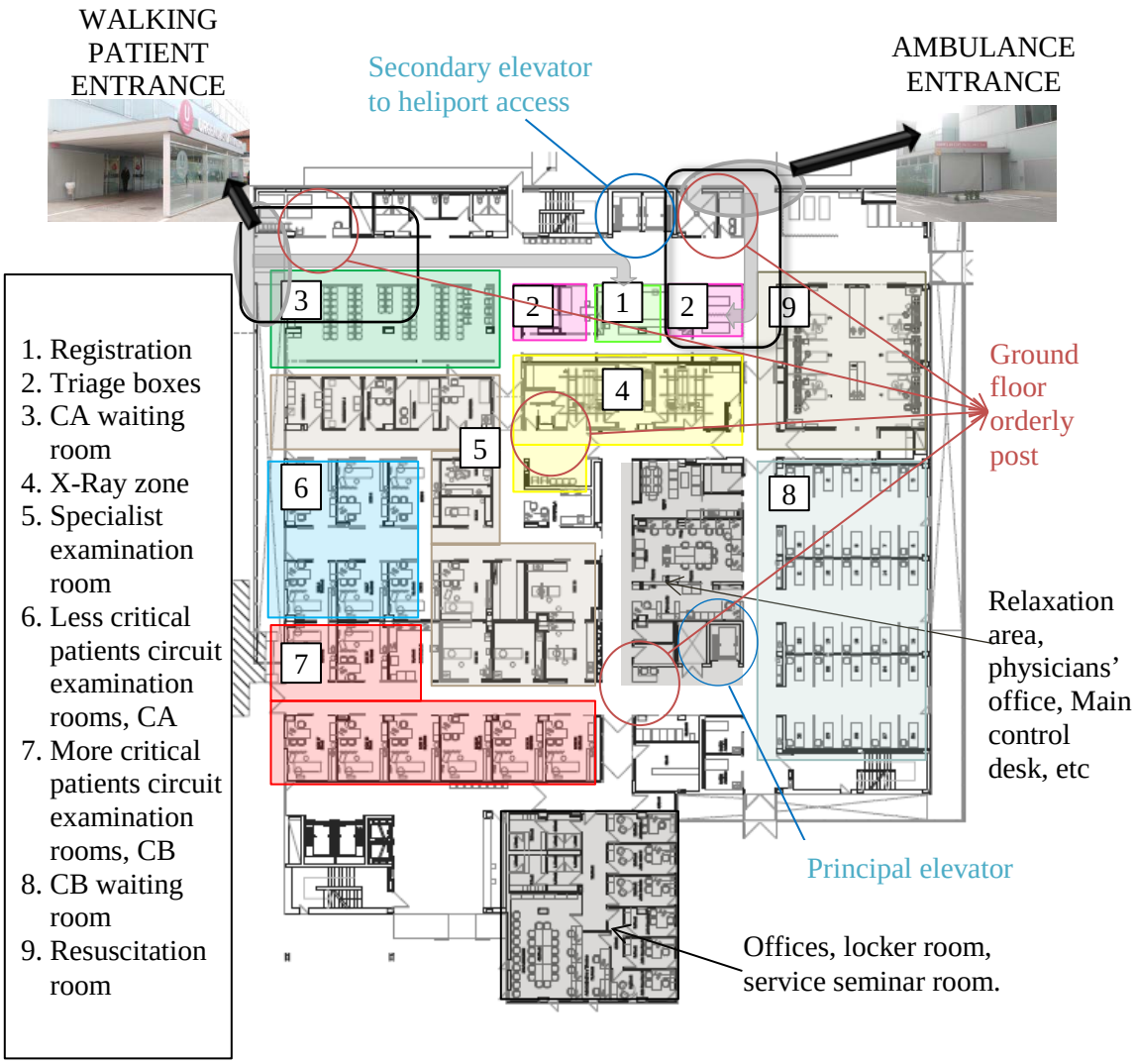


Figure 2.6 HCN ground floor and facilities.

Patient pathways

Analysis of patient records enabled calculation of the probability of a patient following a specific pathway within the ED, together with the associated clinical tests and resources. Each step of the pathway is recorded in the database.

Service time distribution

Although the patient's pathway through the system is automatically registered, the patient record database does not register the time taken for consultations or other processes such as triage, blood tests, X-rays, etc., which had to be recorded in situ in order to obtain estimates.

Estimates were made of the service time for each process, most of which fit a lognormal or a Weibull distribution with different parameter values. Moreover, as board-certified physicians

stated, experience was a relevant factor in the time taken for physicians' consultations, such that first year residents had lower service rates than senior physicians.

2.5.3 Implementation

The construction of the conceptual model and statistical data analysis led to the implementation phase of the study using Arena Simulation ([51], Version 15), a modular simulation software package from Rockwell Automation (see Figure 2.7). Some of the modules include the creation of schedules for different patient arrival rates, calculated in the subsection headed "Patient pattern", different queue management approaches, detailed incident reporting and time spent in each phase of the care process.

The above procedure was coupled with a 3D animation model constructed from the AutoCAD layout of the ED using Sweet Home 3D, sketchup, 3D Max and Arena visual designer. A video of the simulation model can be seen at http://www.unavarra.es/quphs/proyectos, or downloaded specifically at [52].

2.5.4 Verification and validation

Model verification starts with the generated arrivals to the system. This requires comparing the simulated arrival rates for each patient priority group with those calculated from the real historical data. Tests for equality of mean hourly rates and equal variances yielded p-values of >0.05.

Verification of the percentage of patients following the different pathways is carried out by patient priority levels in order to reproduce the same resource requirements as in the real system. Likewise, process time distribution must be verified to ensure accurate computer programming.

Historical data analysis, as well as expert judgments, was used to validate the simulation model and determine its accuracy as a representation of the real ED system and its capacity to generate decisions similar to those that would be made if it were feasible to experiment with the system itself:

- Face validity: the system's experts (ED personnel, managers, specialists, etc.) collaborated by evaluating the model output behavior with respect to that of the real-world system. They interpreted average patient waiting time by priority levels, taking into account care circuit and global system occupational levels, and declared the simulation model a true depiction of reality

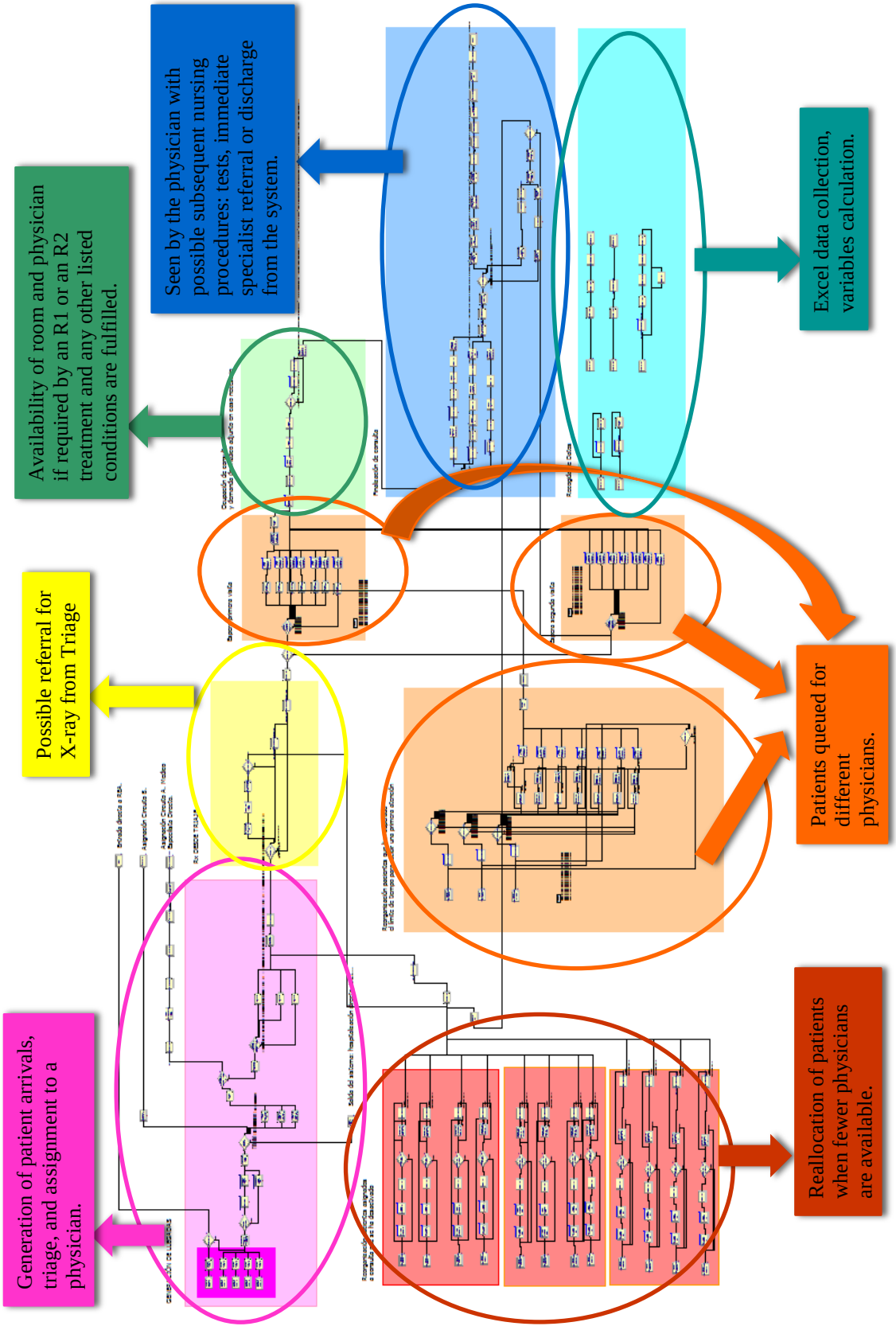


Figure 2.7 Screenshot of the HCN simulation model developed in Arena.

- Animation: medical staff and physicians watched several runs of an animated 3D simulation to check whether its behavior matched that of the real ED. Patients queuing for different ED processes were shown in different colors according to their priority level, to enable viewers to track their movements within the system and the priority rules governing them. Viewers could also check intermediate output values, such as queue lengths and waiting times between processes, and assess their similarity to those experienced in the real-world system.
- Historical data validation: comparison of metrics such as length of stay. Values derived from the system's empirical data were compared with the simulation-derived values using equivalence tests to check the accuracy of the obtained Key Performance Indicators (KPIs)

I. PATIENT FLOW MANAGEMENT

Growth in the utilization of emergency care is observed in high income countries. For instance, emergency admissions grew over 50% from 1992 to 2006 in the US [53] and 9.3% from 2014 to 2017 in England [54], mainly due to the ageing population, which encompasses the main consumers of healthcare services. Some studies quantify that this factor itself can explain 40-50% of the total growth [55], [56]. This trend is expected to continue in the near future. As a result of this growth, the National Center for Health Statistics [57] estimated 43.3 visits to emergency departments (ED) in the US per 100 persons in 2015, which equals a total of 136.9 million visits. Nevertheless, the capacity of healthcare services does not follow the demand growth pace, and it even decreases in some cases [57]. For example, regarding the number of hospital beds per 1000 habitants in the US, which was 4.5 per 1000 in 1980 and 2.5 per 1000 in 2014. Thus, the mixture of a growing demand and a fairly stable capacity of service leads to overcrowded EDs; approximately half of all EDs report operating near or above maximum capacity [58]. This restrictive environment makes operational health care management even more critical, and it is important to guaranty the quality and universality of public healthcare services.

However, EDs are especially difficult to manage; they evolve in a highly stochastic environment due to the variability in the patient arrival rate, illness severity, and, in general, the health resources needed for treatment (material and human) [27]–[29]. In this situation of resource scarcity, the grouping of patients according to their urgency to receive healthcare treatment is a strategy commonly used. Thus, upon arrival, patients undergo an initial assessment, i.e., triage, whose aim is to stratify them by illness acuity and prioritize them accordingly ([59]). Examples of triage systems are the Emergency Severity Index (ESI); the Australasian Triage Scale (ATS), the Manchester Triage Scale (MTS), and the Canadian Triage Acuity Scale (CTAS).

Triage systems may include performance goals in terms of the percentage of patients who should have access to the physician consultation before certain time limits and should have a different time limit and a different percentage for each type of triage-level (see Table 2.2).

Table 2.2. CTAS key performance indicators.

Category	Classification	Access Time	Performance Level
1	Resuscitation	Immediate	98%
2	Emergency	15 min	95%
3	Urgent	30 min	90%
4	Less urgent	60 min	85%
5	Not urgent	120 min	80%

The improvement of the performance of the ED has been addressed by many operations research studies in recent years, such as in Saghafian et al. [15], in which 350 papers dealing with the ED patient flow are reviewed. They distinguish the following three components of the ED patient flow: flow into, within, and out of the ED. The problem addressed in this part of the thesis deals with the patient flow optimization within the ED. In this specific context, Wiler et al. [59] reviews applications concerning patient flow and crowding in the ED. Next subsection explains the patient flow within the ED and its phases.

Patient flow within the ED

Figure 2.8 shows a flowchart of a patient being processed through an ED. Patients arrive either by their own means (normal arrivals) or in an ambulance, and in the first case, the administrative registration process must be carried out. In a very short time, patients access to the examination room, where a triage process classifies the patients according to their severity. Traditionally a nurse evaluates patients at triage, even if there are some papers that have found the investment of a physician at triage to have benefits in the combination of common performance metrics such as LOS, LWBS, and diversion levels ([61]–[65]). In this case the main trade-off from an OR/OM perspective is between 1) using the physician (an expensive resource) at triage instead of in the rooms treating patients, 2) gaining more accurate information upfront, and 3) issuing discharge or appropriate tests early on.

Depending on the hospital and country, the triage process usually uses one of the four ordinal ED triage scales [50] previously mentioned. Without a loss of generality, we consider that the triage classifies ED patients on 5 acuity levels, as is the case of CTAS (Table 2.2: Access time is the upper limit for the arrival to provider time, and performance level is the minimum percentage of patients that should satisfy the access time requirement). Some research considers a complexity-augmented triage proposed by Saghafian et al.[66] would only take a matter of seconds, but it benefits ED performance, both for patient safety and operational efficiency (see for example [67]). Saghafian et al. [66] also investigates several patient flow designs that can be utilized after the complexity-augmented triage is implemented.

After triage, EDs usually organize the patient care into different care circuits; a process known as “streaming”. Usually, a small percentage of total patient volume are subject to a high mortality risk if not treated immediately and they are generally tracked separately from the rest of the patients through a “resuscitation” track. Pioneers in the innovation and implementation

of patients streaming based the streams for the rest of patients on a prediction of their disposition (admit or discharge) made by triage nurses (e.g. [68], [69]). A “fast track” is a stream of dedicated resources used to process lower acuity patients more quickly. Welch [70] notes that a fast track dedicated for minor injuries has been a mainstay in EDs since the 1980s. One of the advantages of this separation is protecting patients with short processing times from waiting behind customers with long processing times.

In the study of this thesis, we will consider the most common structure for EDs. The patient care organized into two different streams or care circuits, apart from the “resuscitation” track dedicated to patients subjected to a high mortality risk if not treated immediately, which constitutes a small percentage of total patient volume. One circuit will be for the treatment of more critical patients and the other for the treatment of less critical patients (“fast track”).

Regardless of the care circuit (stream) where the patients will be treated, once they have been triaged, they wait in a waiting room and are eventually called by a physician. They initially wait in a queue for the first consultation (red arrow in Figure 2.8), in which a physician is needed to evaluate them. This first consultation can result in discharging the patient from the ED (to a hospital ward or to the patient’s home) or in ordering some clinical tests, such as blood tests, X-ray, scan, specialist’s consultation, etc. Once the tests and complementary diagnosis are carried out and their results are ready, the patient re-enters the queue (blue broken line arrow in Figure 2.8) and waits for a second consultation with the ED physician to be reviewed before being discharged from the ED. Note that a patient is usually assigned to a single physician and so must wait for his/her physician to be idle for each consultation.

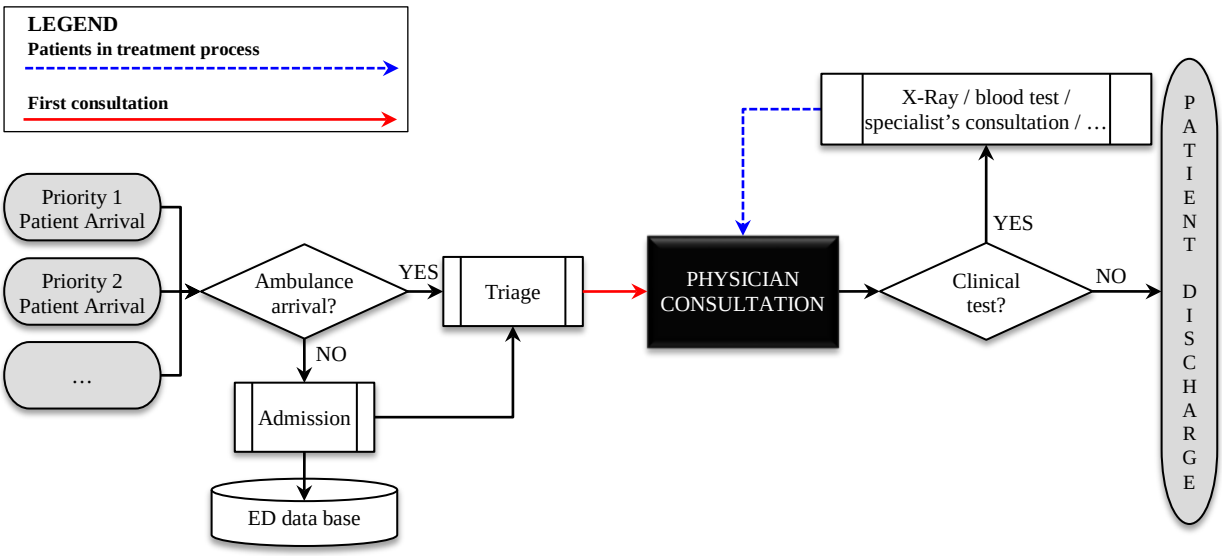


Figure 2.8. Patient flowchart in the ED.

After concluding a consultation, a physician has to choose a pending patient from the queue to provide a medical consultation. This queue is formed by patients of different priorities awaiting initial emergency assessment by the physician or possible reevaluation following tests and/or diagnosis. The queue discipline implemented by the physician greatly influences the quality of the service measures, which is discussed in Chapter 5.

As explained and considered in Saghafian et al. [71], and according to the just explained EDs operation described in Figure 2.8, the flow of patients in the ED is impacted by two phases of sequencing decisions that will be studied in Chapter 4 and Chapter 5 respectively: 1) determination of the order/priority in which patients are initially taken from the waiting area to start treatment with a physician (patient entering the queue represented by a red arrow in Figure 2.8); and 2) determination of the order in which patients are seen once they are under the responsibility of a physician (patient in the queues represented by a red and a blue broken line arrow in Figure 2.8). The phase 1 sequencing decisions are generally made by a nurse considering priority and waiting time in some cases and by physicians in others. The phase 2 sequencing decisions are usually made by individual physicians by choosing among patients assigned to them, who are of different priorities and in different treatment stages. It has been observed wide variance in this latter sequencing logic of individual physicians working within the same ED [71].

With the aim of incorporating a criterion from the physician's point of view to manage the allocation of patients to physicians (first stage of patient flow management), an indicator in real time of the pending instantaneous workload is needed. Both medical and mathematical literature focus on patient's quality of care. Welch et al. [72] and Welch et al. [73] list various metrics by which ED performance can be measured, but there is no metric related to medical staff quality of work or satisfaction to assess performance. Therefore, we address the development of this necessary indicator to manage patient flow optimizing not only traditional performance measures but also physician's quality of work in Chapter 4. We propose an instantaneous workload measure in terms of stress experienced by physicians.

Chapter 3 Workload and stress indicators

3.1 Introduction and related literature

The National Institute for Occupational Safety and Health (NIOSH) defines job stress as the harmful physical and emotional responses that occur when the requirements of the job do not match the capabilities, resources, or needs of the worker [74] . This organization also apprises that the concept of job stress is often confused with challenge, which energizes us psychologically and physically. When a challenge is met, we feel relaxed and satisfied. However, when the challenge turns into job demands that cannot be met the sense of satisfaction turns into feelings of stress.

Job stress level at health services is higher than in other comparable professions [75]. In fact, health care workers have higher rates of substance abuse and suicide than other professions and elevated rates of depression and anxiety linked to job stress [76]. Particularly, EDs are widely known for being a chaotic, stressful, and unpredictable environment within the hospital owing to its stochastic nature. Because of this volatile atmosphere for ED providers, they may be exposed to severe stress most of the time - more than that faced by the physicians of other departments [77]. Specifically, ED physicians of the Hospital Complex of Navarre (HCN) reported experiencing high levels of stress and large inequities regarding the stress from the workload assigned to each of them despite the apparently fair workload assignment rules (e.g. assignment of patients to physicians upon arrival by a simple rotational rule). This statement motivated the research we present in this chapter whose main purpose is, firstly, to provide a method allowing the real-time monitoring of the physician's stress, which is dynamic [78], and secondly, to use it to assess the current HCN-ED physicians stress during the workshift to possibly support their reported feelings of stress and workload inequities among them. This tool will be used in next Chapter 4 to define new patient-physician assignment rules that both reduce and balance the stress among all physicians without worsening other important ED performance measures.

Our cooperation with the ED physicians led us to consider the following stress factors: the workload, the time pressure, and the uncertainty. **Workload** refers to the number and type (severity) of patients that are simultaneously managed by the physician. As patients arrive to the ED, they are triaged (a priority is determined) and immediately assigned to a specific

physician who will be aware of them during their whole health care process. Depending on the hospital and country, this triage process usually uses one out of four ordinal ED triage scales [50]. For example, the Canadian Triage and Acuity Scale (CTAS) classifies patients into five distinct priority levels. Therefore, this workload varies through the work shift because the patient arrivals vary over time and their health status could also evolve over time. **Time pressure** refers to the upper limit for the arrival to provider (ATP) time ("door to doc"), which is defined as the interval between the time a patient arrives at the ED and the time an attending physician sees the patient [73] and depends on the type of patient. Delay in the first diagnostic could put patient health at risk the, especially in those with high severity. Table 2.2 shows the CTAS access time limit as well as the required performance level, which is the minimum percentage of patients that should satisfy the access time limit. The **uncertainty** refers to the lack of knowledge about the patient illness, the tasks needed to provide medical assistance to patients not seen yet or waiting for test results. Generally, the ED healthcare process can be represented by a queue system with several stations, associated to the first and second consultation and some medical tests between them if needed. As the patient is passing through the different stages of their treatment more data about their illness is known reducing the uncertainty.

The considered stress factors are consistent with several studies that also identify the same sources of stress for hospital workers. See for example [79]–[84].

There are several studies that deal with stress measurement methods [85], [86] which can be grouped into two different categories: the first is named *systemic stress*, and it is based on physiology and psychobiology among others (see [87]), and the second is named *psychological stress*, which was developed within the field of cognitive psychology [85], [88]–[90]. In this study, we use "subjective techniques" to quantify immediate physician stress in an ED, particularly the self-report, which belongs to the psychological stress category of measurement. Thereby, physicians provided us with stress assessment data associated to different workload scenarios that are processed by using data analysis. As result, a function able of measuring in real-time the physician's stress is estimated which represents the consensus of the ED physicians about the stress feeling.

Specifically, the stress score associated to a physician workload scenario accounts for the previous mentioned factors: workload assigned to a physician disaggregated by the type of patients (severity), their stage in the medical care process, waiting time targets, and other responsibilities as teaching duties. The proposed methodology involves factor stress analysis, design of questionnaires, and a statistical data analysis of opinions elicited from experts.

Other methods for measuring workload and stress have been proposed in the literature, but none of them can be used to measure and monitor the stress in real time as the one proposed in this chapter does. Well known examples are the Modified Cooper-Harper scale (MCH), the National Aeronautics and Space Administration Task Load Index (TLX), the Overall Workload

(OW) scale, and the Subjective Workload Assessment Technique (SWAT) [91]. These indices provide a global assessment of the total workload for a period of time (e.g. a work-shift). Levin et al. [92] considers "the objective workload" in an ED to be directly proportional to the number of patients being managed simultaneously and inversely proportional to the average severity of them. Specific examples of job stress measures are the Cornell Medical Index [93], the Maslach Burnout Inventory [94], and the Dundee Stress State Questionnaire [95] but they are not suitable to be applied in a dynamic system evolving in a stochastic environment. See more information about job stress measures in [96] review.

The developed stress index can be incorporated into the usual performance criteria measures used in the evaluation of patient flow management policies in an ED. Measures related to patient care are usually used, such as the time until the first consultation, the length of stay or the number of patients in the ED, disregarding indicators related to healthcare staff.

From a production management point of view, patient flow management would be similar to a job shop problem, where the jobs that must be processed are the patients and the different work stations are the different medical consultations and clinical tests. This problem is very important in the production management context and has been the subject of a huge research effort [97]–[99]. However, the problem of managing the flow of patients has its own characteristics, for example that the care pathway of the patients is not known upon their arrival, their health status (and priority) evolves while waiting and the machines of the work stations are mostly people. This specificity of the problem requires the development of management policies specific too.

The methodology of this stress score, which accounts for patients represented as pending workload in Figure 3.1, is also applied to develop a completed workload score, which accounts for discharged patients represented in a red cycle in Figure 3.1. This latter case is not as complicated since workload associated to completely assisted patients is known.

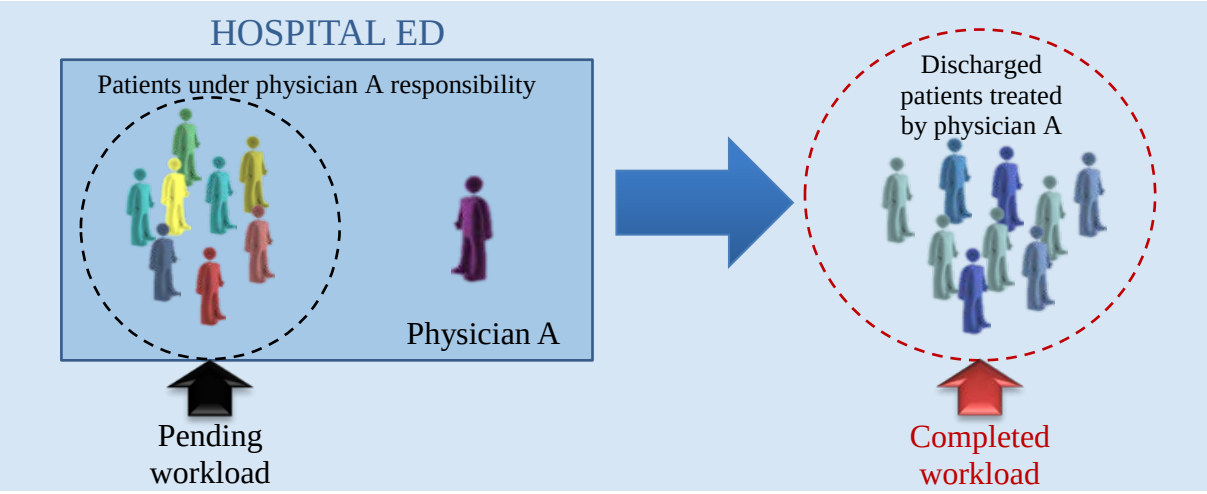


Figure 3.1 Workload considered in the different measures proposed.

This stress score and completed workload, validated and accepted by physicians and that are easily calculated in real-time, serves to monitor the physician's stress and developed workload which can be used to feed a computer application that assigns patients to physicians, as it will be studied in next Chapter 4.

This chapter is organized as follows. Section 3.2 presents all phases necessary to estimate the stress function. In particular, Section 3.2.1 describes step-by-step the preparation for data acquisition, and Section 3.2.2 explains the statistical data analysis. Section 3.3 shows the results of the application of this procedure to a real case in an ED. Analogously, Sections 3.4 and 3.5 presents all phases necessary to estimate the completed workload function and its application to a real case in an ED. Finally, Section 3.6 summarizes and discusses the benefits of our approach.

3.2 Methodology for stress assessment

In this section, we present all steps necessary to estimate a stress function denoted by $f(w)$, whose purpose is to provide a score on a global scale of the stress induced in a physician by a workload scenario w. A workload scenario at a given time t is defined by the set of patients that are currently assigned to a physician. This pending workload includes: patient waiting for the first visit, patients in progress, and patients waiting to be transferred after finishing the medical process in the ED. They change over time: whenever a new patient is assigned to a physician, a new consultation begins or ends, new patient's test arrives, etc.

Let W be the set of all possible different workloads:

$$f : W \longrightarrow R \subseteq \mathbb{R}^+$$

$$w \in W \longrightarrow y_w = f(w)$$

where y_w is the stress induced in a physician when the workload is w, and $R \subseteq \mathbb{R}^+$ is the set of values in which the stress varies.

The aim is to estimate the function $f(w)$ from the statistical analysis of the stress assessment made by experts (physicians working in the ED) from a sample of scenarios representative of W. The methodology is divided in two phases: the first one concerns the preparation for collecting data, in which the job stress factors and their levels are first determined, and then an appropriate survey is designed for eliciting physicians' stress assessments; the second phase covers the data analysis, for which the data is depurated and homogenized, and finally the stress function is estimated. The methodology, structured in seven steps, is summarized in Table 3.1.

Table 3.1. Methodology summary

Phase 1. ***Preparation for data acquisition***. *Step 1.* Identifying the set of factors affecting stress and their categories or levels. *Step 2.* Definition of workload scenarios and selection of a representative sample. *Step 3.* Drawing up the questionnaire to be answered by the experts *Step 4.* Selection of experts, dry run exercise, expert training, and elicitation session.
Phase 2. ***Data analysis*** *Step 5.* Homogenization of experts' answers in a common scale *Step 6.* Table of data. Coherence and consistency analysis for each expert's answers *Step 7.* Estimation of the stress function based on scenario assessments.

3.2.1 Phase 1. Preparation for data acquisition

Step 1. Identifying the set of factors affecting stress and their categories or levels

The stress analysis begins by identifying the set of factors, related to the workload, that affect the physician's stress. Patients, as they arrive to the ED, are triaged and then immediately assigned to a specific physician. Each physician is aware of the pending workload at any moment of the work-shift. The severity level and the waiting time for each patient is known. In addition to the patient consultation work, the physician has to supervise a resident labor during some shifts. All these elements were enumerated by physicians as stressor factors. Table 3.2 represents the factors we consider in our research.

Most EDs have similar structures and ways of operating and consequently similar stressors. However, if the layout of facilities or the ED organization influencing stress are different, the job stress factors summarized in Table 3.2 can be modified and adapted to the particular ED where the methodology is being applied by adding more or substituting them with those job stress factors identified by its physicians.

The stress factors for physicians are grouped into two categories: training responsibility and pending patients. The training responsibility factor refers to the supervision of residents, which are medical school graduates undergoing on-the-job training and cannot assist in all areas of patient demands nor every patient's care needs. Physician can be charged with the supervision of a resident during a whole shift and consequently should have more tasks such as teaching.

All factors, except of the number of patients, are categorical. The factor "resident supervision" has two categories: no resident is supervised and physician supervises a resident.

Table 3.2. Description of different categories for each stress factor.

TRAINING FACTOR	CATEGORIES
RESIDENT SUPERVISION (F_1)	0: No resident supervised
	1: Resident supervised
PENDING PATIENT FACTORS	CATEGORIES
PATIENT PRIORITY (F_2)	1: High priority
	2: Medium priority
	3: low priority
PATIENT MEDICAL ATTENTION PHASE (F_3)	1: Waiting for the C1
	2: In process
	3: Waiting for transfer
PATIENT WAITING TIME TARGETS (F_4)	0: Time limit not exceeded
	1: Time limit exceeded
NUMBER OF PATIENTS (F_5)	Any integer value

Patients in an ED can be of different priorities, which are determined when they are triaged taking into account some medical factors such as the health status, illness, etc. As mentioned in the introduction, these possible priorities depends on the triage scale used by the hospital. In this methodology section we consider an ED where patients can be of priority 1 (high), 2 (medium), or 3 (low). Patients can be waiting for the first consultation (C1), in progress -carrying out medical tests after physician's C1 and waiting for a second consultation (C2)-, or waiting for transfer to their destination (home, hospital) after the medical process in the ED has finished. Thus, "patient medical attention phase" factor has three categories.

Moreover, there are "patient waiting time targets" for the C1, which depends on the patient priority, $t(F_2)$. Two states are considered for this waiting time factor: waiting time below limit and waiting time exceeding the limit. Finally, the factor "number of patients" can take values in the set of all non-negative integers. Table 3.2 summarizes the factors and their categories.

The amount of patients of each type, obtained by combining the levels of the stress factors (F_2, F_3, F_4), are represented by integer variables $X_1, \dots, X_{10}$, while the supervision of residents is coded by a binary variable X_{11} (see Table 3.3).

Table 3.3. Variables originated by the combination of the stress factors.

Variables Description (combination of factors)				Variable Name
Number of pending patients ($F_5 = \sum_{i=1}^{i=10} X_i$)	Medical attention phase(F_3)/Waiting time targets(F_4)		Priority (F_2)	
	1: waiting for the first consultation	1: Time limit exceeded	3	X_1
			4	X_2
			5	X_3
		0: Time limit not exceeded	3	X_4
			4	X_5
			5	X_6
	2: In process		3	X_7
			4	X_8
			5	X_9
	3: Waiting for transfer			X_{10}
Training Responsibility	Resident supervision (F_1)			X_{11}

Step 2. Definition of workload scenarios and selection of a representative sample

We denote by S the workload scenario defined by the variable vector $(X_1, \dots, X_{11})$. For example, $S = (X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 2, X_5 = 0, X_6 = 2, X_7 = 1, X_8 = 3,\ X_9 = 1, X_{10} = 0, X_{11} = 0)$, means that

- There is only a priority 3 patient exceeding the upper limit waiting time ($X_1 = 0, X_2 = 0, X_3 = 1$) and other two priority 1 and two priority 3 patients waiting for the C1 ($X_4 = 2, X_5 = 0, X_6 = 2$).
- There are 5 patients waiting for the C2: one of priority 1, three of priority 2 and one of priority 3 ($X_7 = 1, X_8 = 3,\ X_9 = 1$).
- No patients are waiting for transfer ($X_{10} = 0$)
- No resident supervision ($X_{11} = 0$)

A workload situation w will be represented by a vector S. Because the number of patients assigned to a physician is, theoretically, not capped, the number of different scenarios is also infinite. Furthermore, although the maximum number of patients assigned to a physician was limited by an upper bound, for example fixed according to the maximum value observed in a real ED, the number of different scenarios would also be huge. Figure 3.2 shows the increase of the number of scenarios depending on the maximum number of pending patients. For one pending patient there are 24 different scenarios, but for 15 patients, which is a realistic figure in peak arrival hours, there are over 15 million different scenarios.

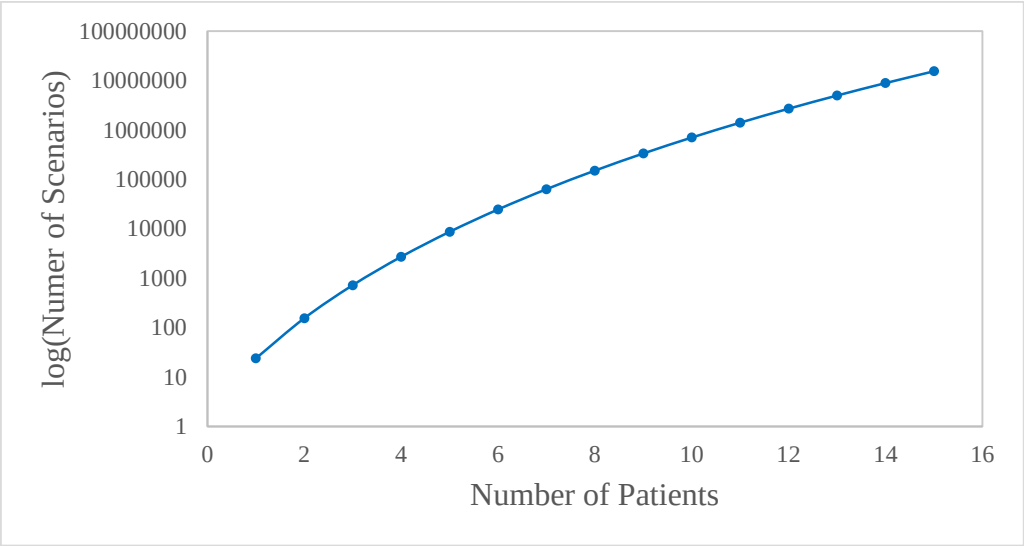


Figure 3.2.Number of possible scenarios depending on the number of pending patients.

Let Ω be the set of possible scenarios $\Omega = \{S_i\}_{i=1}^{\infty}$ and f the stress function:

$$f : \Omega \longrightarrow R \subseteq \mathbb{R}^+$$

$$S \longrightarrow y_S = f(S)$$

where y_S is the stress induced in a physician when the workload w is described by scenario S, and $R \subseteq \mathbb{R}^+$ is the set of values in which the stress varies. Without the loss of generality, we will assume that $R = [0,100]$, with 0 associated to a no stress situation and 100 to a maximum level of stress.

The function f will be estimated from the statistical analysis (see phase 2) of the stress assessment made by physicians working in the ED on a small number of selected scenarios in Ω. The cardinality of Ω prevents an exhaustive assessment of all scenarios S in Ω. To overcome this difficulty, a "D-Optimal" design of experiments, which is a popular criterion that maximizes the determinant of the information matrix, is carried out.

Furthermore, this design has to consider that certain combinations of factor levels may be theoretically possible but very unlikely to be observed in practice. For example, physicians could report that they have never been assigned more than 25 patients. Thus, a set of constraints on the stress variables are imposed on the set of selected scenarios for the design of experiments.

$$P \leq P^* \qquad \sum_{i=1}^{10} X_i \leq L_1 \qquad \sum_{i=1}^{3} X_i \leq L_2 \qquad X_{10} \leq L_3$$

Where $P = \sum_{i=1}^{10} X_i$ is the number of pending patients, P^* is the upper limit of patients assigned to a physician, and $L_j \in \mathbb{N}\ \forall\ j$ is the upper limit for different combinations of patients. This set

of combinations and their associated upper limit values should be suggested by experienced physicians working in the ED. They could vary from one ED to another ED because they depend on the mix of patients attending the ED and other characteristics.

This D-optimal design can be obtained by using the software JMP® [100], which uses an iterative computational method called "coordinate exchange" [101].

The outcome of the design of experiments provided us the necessary scenarios to estimate the main contribution for each factor and first order interactions. Moreover, the constraints defined allowed us to make the factor combinations more realistic and probable to happen through direct contact with medical staff of the ED. This fact makes experts' answers more reliable as they assess more familiar and usual scenarios.

Including extra scenarios as anchors. In many situations, people make estimates by starting from an initial value that is adjusted to yield the final answer, which is biased toward the starting point. This phenomenon is what Tversky and Kahneman call anchoring [102]. In this questionnaire, we will anchor or benchmark experts' answers by defining additional reference scenarios for likely situations in the ED at both ends of the stress scale.

Subjective probability distribution is usually collected from experts by asking them the quantity that corresponds to specified percentiles (usually X_{90} and X_{10}) of his/her subjective probability. In a similar way, as we wanted to rate the scenarios' stress from a range of 0-100, we asked the physician included in our research team to define some realistic scenarios for which the majority of their colleagues would give a very high stress score, U_R, called red scenarios, and others for which the majority of their colleagues would give a very low stress score, U_G, called green scenarios. These green and red scenarios will serve as anchors.

In each set of scenarios given to an expert for stress evaluation there will always be one red or green scenario, which are supposed to be rated at one end of the scale.

The red scenarios include levels of stress factors provided by ED workers that increase the stress feeling (high severity patients, waiting time limits exceeded, large amount of patients to be assisted, etc.). Specifically, the set of red scenarios, Ω_R, are defined as follows:

$$\Omega_R = \{S_j \in \Omega_U \,/\, \nexists\, S_r \in \Omega_U \; satisfying \; S_r < S_j\}$$

where

$$\Omega_U = \{S_j \in \Omega \,/\, f(S_j) \geq U_R\},$$

and U_R is a high level of stress, for example, $U_R \approx 90$.

Similarly, green scenarios were obtained by combining different factor levels provided by ED workers that do not contribute to high levels of stress (low priority patients, waiting time targets achieved, small amount of patients, etc.). The set of green scenarios, Ω_G, are defined as follows:

$$\Omega_G = \{S_j \in \Omega_L \,/\, \nexists\, S_g \in \Omega_L \; satisfying \; S_g > S_j\}$$

where

$$\Omega_L = \{S_j \in \Omega \,/\, f(S_j) \leq U_G\},$$

and U_G is a low level of stress, for example, $U_G \approx 10$.

These extra scenarios, called anchors, augment the number of total scenarios of the design of experiments, while they introduce a maximum for the variability in the range answers. Their utility will be showed in Data Analysis Section.

Step 3. Drawing up the questionnaire for stress assessing

The questionnaire for eliciting the physician opinion concerning the stress associated to a set of scenarios has to be constructed carefully. The design has to simultaneously take into account the difficulty of focus of all potential categories of all the workload factors and for various scenarios. It is known from the work of George A. Miller [103] that there is a limit to our information-processing capacity as the immediate memory span can approximately handle just seven items, and that there is also a span of attention that encompasses a finite number of objects. These considerations lead us to ease the simultaneous processing task required by setting four scenarios for each group that a respondent needs to assess (as shown in the questionnaire included in Appendix B).

In addition, the visual presentation of the scenarios is also important. For example, they feature a native look, just like the physicians' patient portfolio in reality (colour code, structure, etc.). In Figure 3.3, the right-hand side depiction is a capture from the computer screen where a physician consults the pending patients, and the left-hand side represents a scenario as it is included in the questionnaire.

Each scenario shows the list of patients a physician has been assigned. Each patient has a priority (left part of each scenario panel) and is in a specific medical attention phase with possible waiting time targets (colour code). The length of each colour bar indicates the patient priority. Then, there is an indicator in the top right corner that shows if the physician is also supervising a resident (red) or not (white).

Below each scenario, there is an empty box in which the experts have to enter a score based on the stress feeling due to the workload assigned in each patient panel situation represented. Each hypothetical run should be rated on a scale from 0 to 100, where 0 would represent "no stress" and 100 would mean "absolute stress".

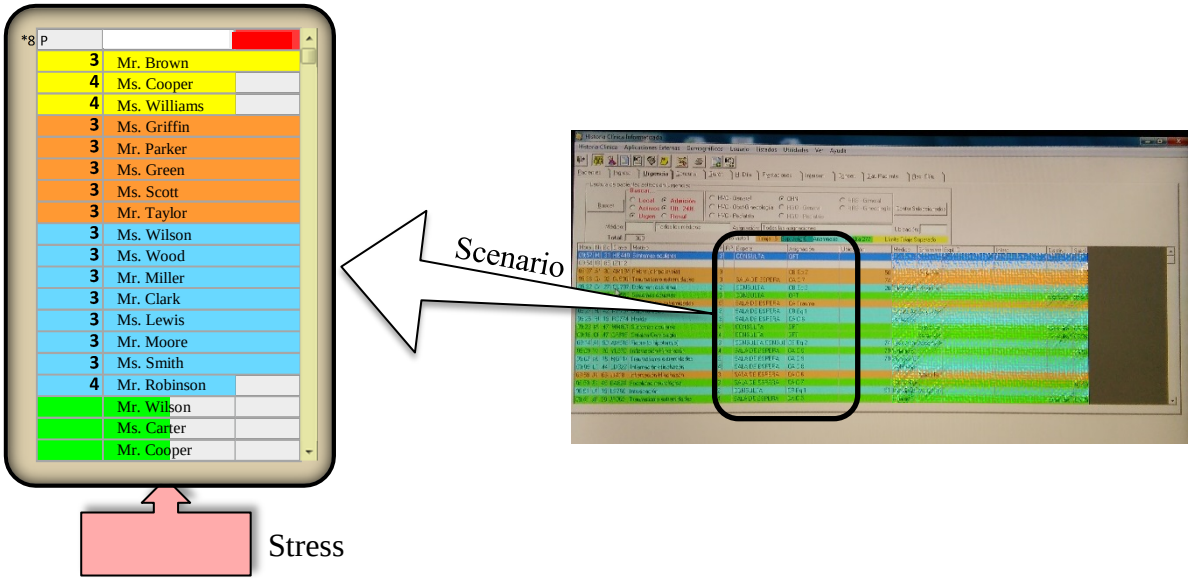


Figure 3.3. Example Scenario of the Questionnaire - Physicians' portfolio of patients in reality

When using the entire scale, it is necessary to give representative numbers for different amounts of workload or stress feeling in examples. Ironson, Smith, Brannick, Gibson, and Paul[38] listed comprehensive evaluative adjective phrases from a survey in the literature in order to orient the respondent's job situation. Similar to Greller and Parsons's [104] effort to develop a psychosomatic measure of work stress, but in terms of workload perceptions, we develop a scale of adjectival items to describe the stressfulness of the job situations (e.g., "no stress", "slight stress", etc.).

To help experts get an idea of the meaning for the quantified stress score, there is a stress scale in the lower half of the questionnaire card with different stress ranges with a qualitative description, as shown in Figure 3.4.

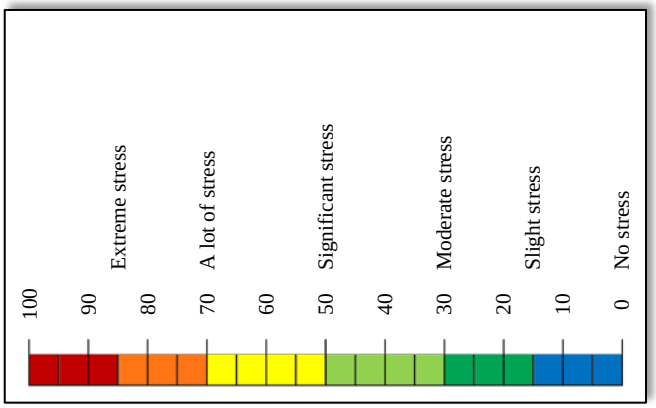


Figure 3.4. Stress qualitative scale.

Each questionnaire consists of four cards, each one with a set of four scenarios such as the one represented in Figure 3.4. That is, each expert is asked to rate the stress in 16 different

scenarios. Each set of scenarios contains an anchor and three other scenarios provided by the D-optimal design of experiments. A number M of different questionnaires are designed in such a way so that no scenario is repeated in a questionnaire, and all scenarios selected by the design of experiments are included throughout the various questionnaires. Furthermore, it is desirable that each scenario is included an equal number of times in the questionnaires. If we have N physicians when we distribute the M different questionnaires among them, some questionnaires will be answered by $\lfloor N/M \rfloor$ physicians and the others by $\lfloor N/M \rfloor + 1$.

Step 4. Selection of experts, dry run exercise, expert training, and elicitation session

The selected experts should be physicians who work in the ED and are accustomed to handling their patients' portfolio in the ED computer screen (whiteboard) - represented in the questionnaire as workload scenarios. They should be familiar in coping with situations similar to those proposed in the questionnaires and have some experience in their work in order to assess them in terms of stress. Furthermore, they should also be interested in the study of stress due to workload (a motivation is to clarify the different stresses produced among physicians because of the patient assignment rule).

First, some experts should be shown the proposed questionnaire to express their remarks and queries. This should help to improve the presentation of the cards and instructions provided in order to make them more clear for a final improved version.

Then, it is necessary to create a training session for the participants in which the objectives of the study and every part of the questionnaire can be clearly explained. It is helpful to provide the experts an instruction sheet - to refer to in case they had doubts while they are filling out the questionnaire - including some guidelines with advice on how to complete the questionnaire and examples to provide familiarity to the way in which the scenarios are represented (see Appendix A and Appendix B).

3.2.2 Phase 2. Data analysis

Step 5. Homogenization of experts' answers in a common scale

One of the problems of general scales is that different raters tend to use different portions and amounts for the scale, which is influenced by personality [105]. In this section, we address the issue of standardizing the opinions of several physicians whose subjective perceptions of stress could differ widely. In other words, different raters may use numbers of a scale in different ways: some experts with a higher threshold for stress may rate all the scenarios in his/her questionnaire, even the most adverse ones, with the maximum score bing 50 in a [0,100] scale, while others tend to crowd themselves into the highest segment of the scale. Meanwhile, there could also be experts who spread their score values across the whole scale.

This issue is addressed by a mathematical transformation of the physician's scores in order to spread them all over the scale range. In this way, the transformed scores from different physicians are comparable.

Let $y_i^*(S_j) = y_{ij}^*$ be the stress score for scenario S_j provided by physician i. In case the range of values $\{y_{ij}^* / j = 1, \dots, 16\}$ greatly differs from the total range [0, 100], a transformation $g(y_{ij}^*) = y_{ij}$ is needed such that the range of values $\{y_{ij} / j = 1, \dots, 16\}$ is similar to [0, 100].

This transformation should preserve the ordering of scenarios and the ratio of differences in stress among them. Any non-decreasing transformation preserves the ordering of the scenarios. In addition, the second condition lead us to a linear transformation.

That is, for any given two scenarios S_u, S_v and any scenario S_j, it is imposed that

$$\frac{y_{ij}^* - y_{iv}^*}{y_{iu}^* - y_{iv}^*} = \frac{y_{ij} - y_{iv}}{y_{iu} - y_{iv}}$$

From which a linear relationship between y_{ij} and y_{ij}^* is readily obtained:

$$\boldsymbol{y_{ij}} = \frac{y_{iu} - y_{iv}}{y_{iu}^* - y_{iv}^*} \times \boldsymbol{y_{ij}^*} + \left(y_{iv} - \frac{y_{iu} - y_{iv}}{y_{iu}^* - y_{iv}^*} \times y_{iv}^* \right)$$

$$y_{ij} = g(y_{ij}^*) = a_i + b_i \times y_{ij}^*$$

Where

$$a_i = \left(y_{iv} - \frac{y_{iu} - y_{iv}}{y_{iu}^* - y_{iv}^*} \times y_{iv}^* \right)$$

$$b_i = \left(\frac{y_{iu} - y_{iv}}{y_{iu}^* - y_{iv}^*} \right)$$

The scenarios providing the pairs (y_{iu}^*, y_{iu}), (y_{iv}^*, y_{iv}), which determine the parameters of the linear transformation, are those introduced as anchors in the questionnaire. U_R is the expected stress induced by a red scenario. This value is estimated by the trimmed mean of the scores provided by physicians for those red scenarios:

$$y_i^R = \max\{y_{ir} / S_r \in \Omega_R\}$$

$$\{y_{[j]}^R\}_{j=1}^N \textit{ the ascending ordered set } \{y_i^R\}_{i=1}^N$$

$$\frac{1}{n - 2p} \sum_{j=p+1}^{n-p} y_{[j]}^R = \widehat{U}_R$$

Furthermore, the confidence interval (CI) for U_R is calculated, and those physicians, whose scores for red scenarios are below the left limit and gave scores on the lower side of the scale – e.g. because they have a higher stress threshold than their colleges – need to be rescaled according to the linear transformation.

U_G is the expected stress induced by a green scenario, and its value is estimated by the trimmed mean of the scores provided by physicians for those green scenarios $\widehat{U}_G$, similar to $\widehat{U}_R$. The CI for U_G is also calculated and those physicians, whose scores for green scenarios are above the right limit, need to be rescaled.

When a physician i with a high stress threshold uses only the low side of the stress scale and their scores for red scenarios, y_i^R, are below the CI calculated (case 1), then the pairs for the transformation are $\left(y_i^R, \widehat{U}_R\right)$ and $\left(y_i^G, y_i^G\right)$. In the opposite case, when a physician i has all their scenario scores in the upper side of the stress scale (case 2) and y_i^G is above the CI calculated for U_G, the pairs for the transformation are (y_i^R, y_i^R) and $(y_i^G, \widehat{U}_G)$. Finally, if a physician i has all their values concentrated on the middle of the scale (case 3), then the pairs for the linear transformation are $\left(y_i^R, \widehat{U}_R\right)$ and $(y_i^G, \widehat{U}_R)$.

As mentioned, if a physician i has spread all their values over the stress scale (case 4), these do not need to be rescaled. These four homogenization cases are represented in Figure 3.5, where PS scores are the scores provided by physicians on their personal scale, and CS scores are the physicians' scores on the common scale.

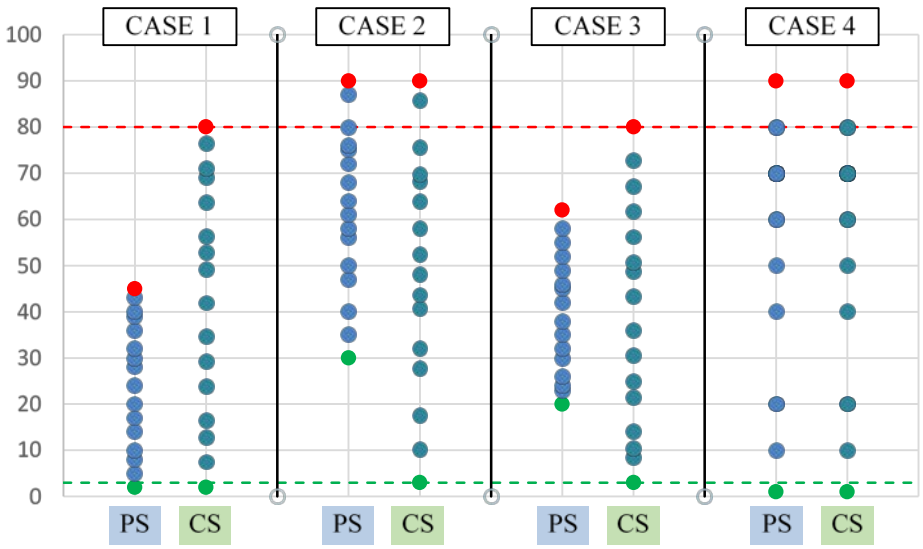


Figure 3.5. Cases of homogenization for physicians' scores.

Step 6. Table of data. Coherence and consistency analysis for each expert's answers

In this section, the internal coherence and consistency among raters are analysed. Stress scores coming from no-coherent or inconsistent physicians have to be discarded.

Coherence. To analyse the coherence of a physician, it is necessary to introduce the concept of dominance between scenarios:

A scenario S_1 defined by the vector of stress variables $\{X_{1,1}, \dots, X_{1,11}\}$ dominates over a scenario $S_2 \equiv \{X_{2,1}, \dots, X_{2,11}\}$, represented by $S_1 > S_2$, if and only if

$$X_{1k} \geq X_{2k} \; \forall \; k = 1, \dots, 11 \; and \; \exists \; j \; s.t. X_{1j} > X_{2j}$$

A physician is coherent assessing scenario S_1 and S_2 if $S_1 > S_2 \Leftrightarrow y_i(S_1) > y_j(S_2)$.

A coherence index, CoI, similar to the Kendall's tau-a is defined by taking into account the pairs of scenarios with a dominance relationship which is coherently and incoherently assessed, denoted by D_c and D_D, respectively:

$$CoI = \frac{D_c - D_D}{D_c + D_D}$$

Physician whose CoI are below a certain threshold are excluded.

Consistency. Many researchers in medicine, biology, engineering, etc. need measures of agreement aimed to assess the reproducibility of judgements. The concept of inter-rater reliability expresses our need of quality for measurement, in terms of concordance of judgments - as this study looks for a consensus among physician. The assessment of this inter-rater agreement has been extensively studied [106]–[112] are examples of these studies).

Most of them propose the Kappa Statistic, a statistic that indicates the degree of agreement from nominal or ordinal assessments. However, when there are ordinal ratings, Kendall's coefficients are more appropriate statistics to determine association as they take ordering into consideration.

We check the consistency of a physician by comparing his/her answers with those physicians that answered the same questionnaire. Thus, "Kendall's correlation coefficient tau-b" could be more appropriate to use as it measures association between two ordinal variables, each appraiser (one physician) with the known "standard" (the consensus from rest of the group).

Now, the question of how to define the "standard" arises. One possibility is to create the standard by averaging the scores provided by other physicians or by selecting the median

answer or other statistic. However, these values could not represent the majority's opinion of the group. Suppose three raters provide (10, 10, 15) to S_1 and (12, 12, 6) to S_2, the majority agrees that $S_2$is more stressful than S_1, but $\overline{y(S_1)} < \overline{y(S_2)}$. To avoid these undesirable situations, we define the standard directly from a voting system. One scenario is considered more stressful than the other when the majority of the group considers it so. If there is a tie, then we have a "indecisiveness" situation. The Kendall's tau-b is adapted to consider the three possible cases of concordance (C), discordance (D), and indecisiveness (I), and a new index is defined:

$$CGI = \frac{(C - D)}{(C + D + I)}$$

More details about the calculation for this index are explained in Appendix C.

Physicians whose CGI are below a certain threshold are excluded.

Step 7. Estimation of the stress function based on scenarios assessments.

Once physicians' scores for scenarios have been rescaled when necessary, and coherence and consistency controls have been carried out, the stress function is estimated by regression techniques.

The rescaled – when necessary - stress felt by a physician i, $\Upsilon_{\mathrm{i}}(S_j) = \Upsilon_{\mathrm{ij}}$, when the scenario S_j represents the instantaneous workload assigned, can be expressed as

$$\Upsilon_{\mathrm{ij}} = f(S_j) + \mathcal{E}_{ij}$$

where $f(S_j)$ is the stress induced by the scenario S_j, which could be interpreted as the consensus score [113], true score [114], or universal score [115] for the workload of S_j over the population. The residual $\mathcal{E}_{ij}$ carries the unique effect for physician i. This personal component, $\mathcal{E}_{ij}$, is due to the person's reality perception, personality, years of experience, capability, etc. It is assumed that, $E(\Upsilon_{\mathrm{ij}}) = f(S_j)$, and then $E(\mathcal{E}_{ij}) = 0$.

In order to keep the stress scores in the range [0, 100], a multiple linear regression with a logit link for the stress score is proposed. The independent variables are the stress variables $\boldsymbol{X} = \{X_1, \dots, X_{11}\}$.

$$logit(Y(S)) = \log\left(\frac{Y(S)}{100 - Y(S)}\right) = g(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{11} X_{11} + \varepsilon$$

$$Y(S) = 100 \times \frac{e^{f(\boldsymbol{X})}}{1 + e^{f(\boldsymbol{X})}}$$

3.3 Stress assessment: a case study

This proposed methodology was applied to analyse the stress of physicians in the ED of the HCN, which is staffed 24 hours per day with 43 board-certified emergency physicians and is described in detail in Chapter 2.

3.3.1 Phase 1. Preparation for data acquisition

Identifying the set of factors and their categories. Definition of workload scenarios and selection of a representative sample. In the first step, we widely discussed with the ED physician staff in order to define every factor affecting stress, as well as all their possible combinations to pinpoint the stress variables. Then, we created the design of experiments imposing some constraints provided by the physician in my research team, who has more than 10 years of experience, and was essential in the application of all steps in Phase 1. These constraints allowed us to construct more realistic scenarios that had a high probability of happening in the ED based on expert raters, e.g., no more than 19 patients being managed simultaneously ($P \leq 16$), no more than eight patients waiting for the first consultation simultaneously ($\sum_{i=7}^{9} X_i \leq 8$), no more than two patients of priority 3 waiting for the first consultation with the time limit exceeded ($X_1 \leq 2$), etc.

The outcome of the "D-optimal" design carried out provided us 72 different scenarios to assess. These were the necessary scenarios to estimate the main contribution of each factor and first order interactions. Furthermore, we designed 12 extra scenarios as anchors – six for high stress levels and six for low stress levels. That is, from a panel of experts, we were able to obtain scores for a total of 84 scenarios. We designed six different questionnaires containing four cards with four different scenarios on each one (16 in total, see Appendix B).

Drawing up the questionnaire for stress assessing. As mentioned in Section 3.2.1, the scenarios of the questionnaire were designed to feature a native look to be easily associated to real workload situations in the ED department. They imitate the ED physicians' computer screen (see Figure 3.3 and Appendix B).

To form the questionnaire, all scenarios were originally ordered with estimated guided values of the factors' coefficients previously provided by the perception from a couple of the ED-physicians and our research team's physician. We grouped the 72 scenarios in 12 stress ranges –three main stress ranges subdivided into four sub-ranges – and assigned scenarios of all 12 ranges in each questionnaire, and specified one scenario for each main stress range to each questionnaire card.

Finally, we randomly assigned two green and two red scenarios to each questionnaire (one for each card) to augment variability with extreme scenarios on both sides of the scale and anchor the answers, as we have explained in Section 3.2. That is, each set of scenarios contained an

anchor, and three other scenarios were randomly selected from the different ranges of the stress scale.

Selection of experts, dry run exercise, expert training, and elicitation session. The questionnaire was presented and provided to all physicians of the ED (43 physicians) in a training session (see Appendix B). In the training session, we discussed every part of the questionnaire, showed some example responses, and gave them an instruction sheet to refer to as a guide. After two weeks –with a reminder in the middle of that period- we got the 70% of the ED physicians staff to answer the questionnaire. The final panel was made up of a total of 30 ED physicians: 13 with more than 15 years of experience, 11 with 5-10 years of experience, three with less than five years of experience, and three with unknown experience. They found the questionnaire reasonable and the scenarios very similar to their usual work situations.

We finally collected a total of 472 stress scores for the 84 scenarios (there were two instances where last card of the questionnaire was overlooked - four scenarios on each one), and each scenario was rated by a minimum of four and a maximum of six different physicians.

3.3.2 Phase 2. Data analysis

Homogenization of stress scales. We calculated the stress score for green and red scenarios (both extreme sides of the stress scale), and we obtained three and 80 as the lower and the upper limit of the common scale range, respectively. Based on these values, we only rescaled the physicians' opinions whose minimum score was over the minimum limit or the maximum below the upper limit.

Coherence and consistency of raters. We first analysed each physician's response in order to detect "incoherent" experts. They all scored with a higher stress value in the scenarios which dominate others, so we could not discard any physicians due to his/her incoherence (CoI=1).

Then, we calculated the consistency with the group index, CGI and the Kendall's tau-b consistency index. The answers of these experts were carefully checked, which revealed physicians inconsistent with his/her group who significantly ordered his/her questionnaire's scenarios differently in terms of stress. As the purpose of the study was to assess the stress of ED physicians in the workplace due to workload, and the results should be validated by the experts and represent a consensus among them, we didn't take into account his/her questionnaire in order to improve the stress assessment accuracy.

Figure 3.5 shows the results of the group with the inconsistent physician (11), which has a low CGI value:

Table 3.4. Consistence of physicians belonging to Group 3.

	Physician	Kendall's tau-B	p-value	GCI
Group 3	11	0.38	6.40E-02	0.07
	12	0.69	3.29E-04	0.74
	13	0.63	1.46E-03	0.58
	14	0.70	3.97E-04	0.50
	15	0.80	3.54E-05	0.61

However, the rest of the groups did not present inconsistencies, and they generally agreed in their stress scores and scenarios' order. As an example, Figure 3.6 shows the score that different physicians gave to the same scenarios (questionnaire 2 in this case). All physicians have a similar consistency with the group index.

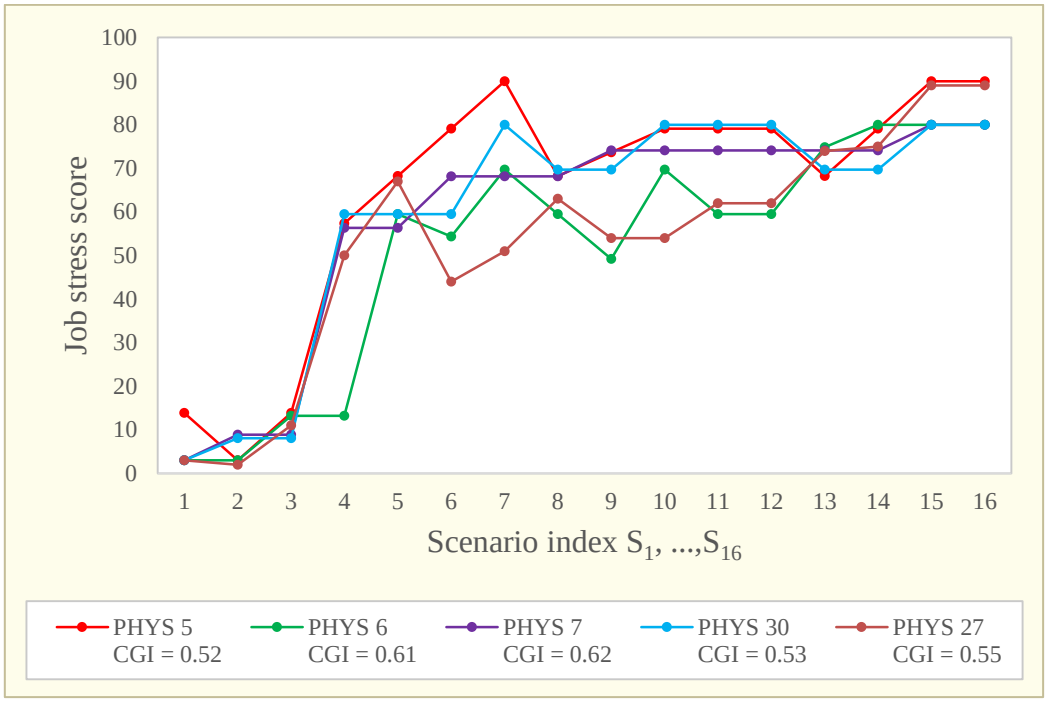


Figure 3.6. Group 2 scenarios' score.

Estimation of the stress function based on scenarios assessments. The data from the rest of the physicians (29) was taken into consideration to run the multiple linear regression with the logit of the stress score as the dependent variable (see Appendix C). We obtained that all types of patients were statistically significant for the dependent variable. We discovered that those patients who have received the complete care process but are waiting to be transferred have less of an effect than the rest of patient scenarios (p-value=0.05).

However, supervision training, which had been mentioned many times by physicians as one of the most important factors for stress, has a higher p-value (0.238) and are statistically not as

significant as the rest of factors, but we considered all 11 variables (included the training supervision) for the model.

Table 3.5. Regression coefficients

Model variables (combining stress factors)		Coef.	p-value
Number of patients in each care situation			
Waiting for the first consultation ($F_3 = 1$)			
Time limit exceeded ($F_4 = 1$)			
Number of Priority 3 patients	X_1	0.726	0.000
Number of Priority 4 patients	X_2	0.458	0.000
Number of Priority 5 patients	X_3	0.410	0.000
Time limit not exceeded ($F_4 = 0$)			
Number of Priority 3 patients	X_4	0.313	0.000
Number of Priority 4 patients	X_5	0.279	0.000
Number of Priority 5 patients	X_6	0.207	0.000
In process ($F_3 = 2$)			
Number of Priority 3 patients	X_7	0.189	0.000
Number of Priority 4 patients	X_8	0.155	0.000
Number of Priority 5 patients	X_9	0.182	0.000
Waiting for transfer ($F_3 = 3$)			
Number of patients	X_{10}	0.113	0.005
Training supervision			
Resident supervision	X_{11}	0.078	0.238

The chosen model yielded a determination coefficient of above 0.70, and the regression function was the following:

$$Y(S_j) = Y_j = 100 \times \frac{e^{f(\boldsymbol{X})}}{1 + e^{f(\boldsymbol{X})}}$$

$$f(\boldsymbol{X}) = -3.378 + 0.726X_1 + 0.458X_2 + 0.410X_3 + 0.313X_4 + 0.280X_5 + 0.207X_6 + 0.189X_7 + 0.155X_8 + 0.182X_9 + 0.113X_{10} + 0.0778X_{11}$$

$$X_1, \dots, X_{10} \in \mathbb{N}$$

$$X_{11} \in \{0,1\}$$

This model allows us to assess every possible situation in the ED through the workload information of the physicians' whiteboard (patients assigned). Figure 3.7 shows the stress associated to different workload scenarios ordered from least stressful to most stressful. There are four scenarios, which belong to the 84 designed scenarios for the questionnaires, represented on the figure.

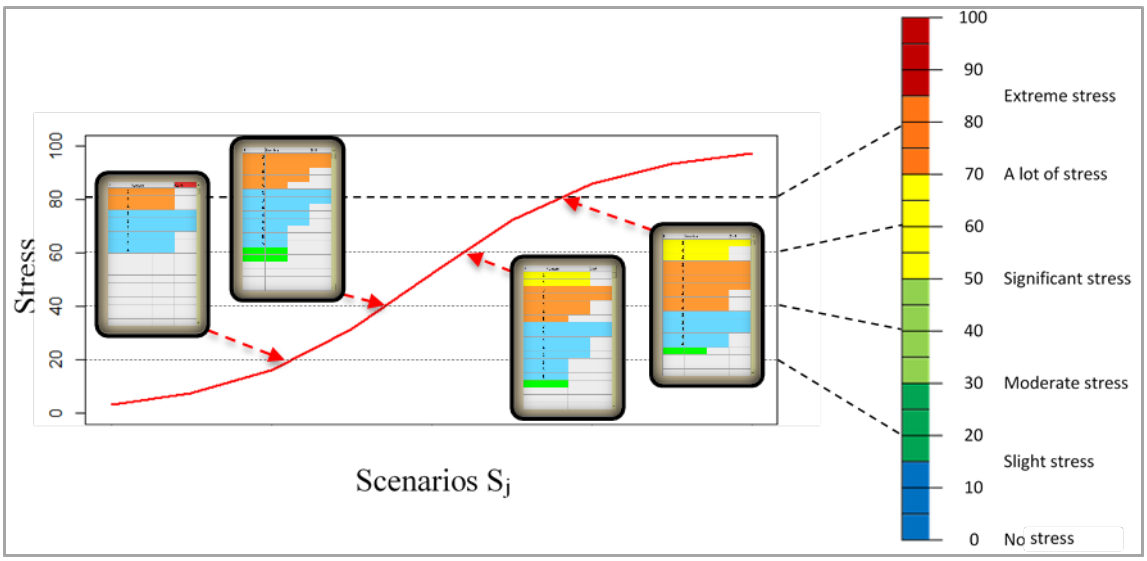


Figure 3.7. ED-situations stress assessment

This model was validated both statistically and by physicians. All statistical assumptions for this regression were met. For example, Figure 3.8 shows that residuals are distributed normally and have a mean of zero.

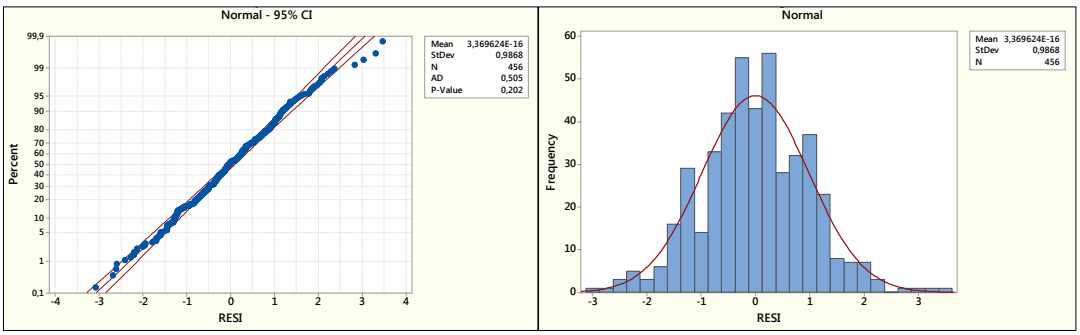


Figure 3.8. Probability Plot and Histogram of residuals

Model validity was also checked by the ED physicians, who were asked to test the results and gave their approval. Table 3.5 clearly shows the factor's influence on physician stress through the variable's coefficients. For the factors of the type of patients, F_3 and F_4 are represented with the color code used in the ED, while F_1 (priority) is in a descending order (inverse to severity index).

Patients, who have not yet been seen by a physician for the first time – and consequently, the physician could not have known the health condition or care needs, nor could have provided medical treatment, request medical tests etc. – have the highest coefficient ($X_1 - X_6$). Within that group, patients whose waiting time has been exceeded ($X_1 - X_3$), represented in yellow (see Table 3.5), contribute to higher levels of stress than the others ($X_4 - X_6$), represented in orange. This result supports theories that state that uncertainty and time pressure are some of the most prevalent causes of anxiety, which is a symptom of stress [116]–[118].

As patients have less process stages left ($(X_1 - X_6) > (X_7 - X_9) > (X_{10})$), their contribution to the physicians' overall stress decreases. Moreover, within all these groups of patients that produce a high amount of uncertainty and time pressure, the most severe a patient is the more stress he/she logically produces for the physicians.

3.4 Methodology for workload assessment in a shift

In this section, we present all steps necessary to estimate a workload function denoted by $C(p)$, whose purpose is to provide a score of the effort made by a physician when p is the workload completed. The workload completed from the beginning of their work shift until a given time t by a physician is defined as the result of the sum of work associated of each patient completely assisted and discharged from the system. This workload is the servers capacity actually required to perform the patients' assistance, which includes the time needed, physician intensity of the work and mental effort required. It is considered to be a primary source of resource depletion. They change over time: whenever a new patient is discharged. A patient who has just been discharged ends up inducing stress on the physician that has been responsible for their complete care (previous measure developed in Section 3.2). Thus, the patient becomes part of the workload completed function.

Let P be the set of all possible different workload completed:

$$C: P \longrightarrow R \subseteq \mathbb{R}^+$$

$$p \in P \longrightarrow z_p = C(p)$$

where y_p is the effort made by a physician when p is the workload completed, and $R \subseteq \mathbb{R}^+$ is the set of values in which the effort varies. It is a function of the number of patients completely assisted of each type.

The aim is to estimate the function $C(p)$ from the statistical analysis of the assessment of the workload associated to the complete assistance of the different type of patients made by experts (physicians working in the ED) by using the Analytic Hierarchy Process (AHP). It is a theory of measurement first developed within the management science field by Saaty in 1980 [119] through pairwise comparisons and relies on the judgements of experts to derive priority scales. These comparisons are made using a scale of absolute judgements that represents, how much more, one element dominates another with respect to a given attribute (workload associated).

As detailed explained in Section 3.2, the methodology, which is summarized in Table 3.6 is divided in two phases: the preparation for collecting data and the data analysis.

Table 3.6. Methodology summary

Phase 1. ***Preparation for data acquisition***. *Step 1.* Identifying the set of factors affecting workload and their categories or levels. *Step 2.* Definition of completed workload scenarios. *Step 3.* Drawing up the questionnaire for workload assessing. *Step 4.* Selection of experts, dry run exercise, expert training, and elicitation session.
Phase 2. ***Data analysis*** *Step 5.* Table of data. Consistency analysis for each expert's answers. *Step 6.* Aggregation of expert's answers. *Step 7.* Estimation of the workload function based on pairwise comparison.

3.4.1 Phase 1. Preparation for data acquisition

Step 1. Identifying the set of factors affecting workloads and their categories or levels

The workload analysis begins by identifying the set of factors that influences the servers capacity actually required to perform the patients' assistance, which includes the time needed, physician intensity of the work and mental effort required. As patients are triaged, they are assigned to a physician who is responsible for their treatment until they are discharged from the ED. Each physician has a pending workload at any moment of the work-shift that makes them experience some stress. Once a patient is discharged, they do not take part in the pending workload anymore and become part of the workload completed by their responsible physician. At this moment, the severity level as well as their care needs are known. Both elements were enumerated by physicians as workload factors and are represented in Table 3.7.

Most EDs have similar structures and ways of operating and consequently similar patient workload factors. However, as in the previous methodology, if the patients' care needs or the ED organization influencing workload are different, the workload factors summarized in Table 3.7 can be modified and adapted to the particular ED.

All factors influencing patients' workload for physicians are categorical. The factor "patient care needs has two categories: "a single consultation" and "more than one consultation". These refer to patients that needed a single consultation, and those who needed a first consultation, medical tests requests, and a second consultation to be reevaluated before being discharged.

Table 3.7. Description of different categories for each workload completed factor.

DISCHARGED PATIENT WORKLAOD FACTORS	CATEGORIES
PATIENT PRIORITY (F_1)	1: High priority
	2: Medium priority
	3: low priority
PATIENT CARE NEEDS (F_2)	1: A single consultation
	2: More than one consultation

As in previous sections (Section 3.2 and 3.3), we also consider in the methodology an ED where patients can be of priority 1 (high), 2 (medium), or 3 (low), which is determined when they are triaged. All patients considered have already been discharged from the ED as their medical care in the system has finished.

The amount of patients type workload are obtained by combining the levels of the workload factors (F_1, F_2) and are represented by variables $\theta_1, \dots, \theta_6$ (see Table 3.8).

Table 3.8. Variables originated by the combination of the stress factors.

Variables Description (combination of factors)		Variable Name
Care needs (F_2)	Priority (F_1)	
1: A single consultation	3	θ_1
	4	θ_2
	5	θ_3
2: More than one consultation	3	θ_4
	4	θ_5
	5	θ_6

Step 2. Definition of completed workload scenarios

We denote by S the workload completed scenario defined by the integer variable vector $(n_1, \dots, n_6)$, which are the number of patients of each type discharged by a physician. For example, $S = (n_1 = 1, n_2 = 0, n_3 = 2, n_4 = 0, n_5 = 2, n_6 = 1)$, means that

- There is a priority 3 and two priority 5 patients that have been discharged after the first consultation $(n_1 = 1, n_2 = 0, n_3 = 2)$.
- There are 2 patients that have been discharged after needing a first consultation, have some medical tests done and a second consultation when their results were ready to be reevaluated before being discharged.

A workload complete situation p will be represented by a vector S. Because the number of patients assisted to a physician is, theoretically, not capped, the number of different scenarios is also infinite. Furthermore, although the maximum number of patients assigned to a physician

was limited by an upper bound, for example fixed according to the maximum value observed in a real ED, the number of different scenarios would also be huge. However, we consider that the depletion of physicians due to every task is accumulative and workload due to every patient completely assisted should be added in order to calculate the total workload completed from the beginning of the shift to any time t.

Let Φ be the set of possible scenarios $\Phi = \{S_i\}_{i=1}^{\infty}$ and c the completed workload function:

$$c : \Phi \longrightarrow R \subseteq \mathbb{R}^+$$

$$S \longrightarrow z_S = c(S)$$

where z_S is the labor completed by the physician when the completed workloads (patient's completely assisted) p are described by scenario S, and $R \subseteq \mathbb{R}^+$ is the set of values in which the workload varies.

The function c,

$$c(S) = \sum_{i\,\epsilon} n_i\,\theta_i$$

will be estimated from the statistical analysis (see phase 2, Section 3.4.2) of the workload $(\theta_1, \dots, \theta_6)$ assessment made by physicians working in the ED.

Step 3. Drawing up the questionnaire for workload assessing

In this section it will be described how to elicit experts' opinion about the workload associated to each patient care to estimate $\theta_1, \dots, \theta_6$.

The assignment of ordinals to different previously defined tasks preserves order but carry no information about differences or ratios of relative magnitudes so it is crucial to use scales of measurement. They consists of three elements: a set of objects, a set of numbers, and a mapping of the objects to the numbers. Particularly, a standard scale can be used to measure objects or events with respect to the property for which a scale is designed to measure. Since the unit is arbitrary, one can have different numbers to which the objects are mapped. However, we must be constantly and carefully attentive to how we interpret data from scales and note that standard scales force on us a to think in way that is not in complete harmony with the way we really feel about what they are measuring.

As mention in the introduction of this section, we propose the use of the AHP in which a more general method of measurement is used: the method of relative measurements, which is useful for properties for which there is no standard scale of measurement (intangible properties). Moreover, measurements in a standard ratio scale are transformed to measurements in a relative ratio scale by normalizing them. The AHP method helps us to derive relative scales using

judgment or data from a standard scale, and how to perform the subsequent arithmetic operation on such scales avoiding useless number crunching.

The judgments are given in the form of paired comparisons ([120], [121]) as the most effective way to concentrate judgement is to take a pair of elements and compare them on a single property without concern for other properties or other elements. We also note that sometimes comparisons are made on the basis of standards established in memory through experience or training.

In this study we have only one criteria to evaluate the 6 activities, workload associated to them. Thus, we need to construct a 6x6 matrix whose entries reflect the relative workload of one element compared to the others. The values are usually in the interval of 1–9 or their reciprocals. As the matrix is a square reciprocal matrix, $A = [a_{ij}]_{n\times n}$, $a_{ij} = \frac{1}{a_{ji}}$ and the questionnaire contains 15 subjective pairwise comparisons such as the following, in which θ_1 is compared to θ_5:

	PATIENT TYPE	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	PATIENT TYPE
8	θ_1 ($F_1 = 3; F_2 = 2$) *P3, consultation + + medical tests + + consultation.*																		θ_5 ($F_1 = 5; F_2 = 1$) *P5, single consultation.*

Figure 3.9. Example of pairwise comparison elicitation in the questionnaire.

In previous comparison (Figure 3.9), expert should mark how greater the workload associated to the complete assistance of a type θ_1 patient (patient of high priority who had needed one consultation, some medical test, and a second consultation to be reevaluated before being discharged) is compared to that associated to the other patient's θ_5 (priority 5 patient who had been discharged after a single first consultation), on a scale of 1 to 9. This recommended by Satty (1980) scale of relative importance from 1 to 9 for making subjective pairwise comparisons is adapted in Table 3.9 and will be provided with instructions and questionnaires to experts participating in the study to help them to get an idea of the meaning for the values.

Table 3.9. Verbal judgement: 9-Point intensity or relative importance scale.

Scale	Definition	Explanation
1	Equal importance	The complete assistance of both types of patients require **equal** workload.
3	Moderate importance of one over the other	The complete assistance of the patient type closest to the box implies **moderately greater** workload than the other's.
5	Essential or strong importance	The complete assistance of the patient type closest to the box implies **essentially greater** workload than the other's.
7	Very strong importance	The complete assistance of the patient type closest to the box implies **much greater** workload than the other's.
9	Extreme importance	The complete assistance of the patient type closest to the box implies **extremely greater** workload than the other's.
2,4,6,8	Intermediate values between the two adjacent judgments	When compromise is needed

Figure 3.10 is an example of response to the pairwise comparison described in Figure 3.9. The type of patient who is closer to the marked box is the dominant, that is, the patient with the most workload associated. The interpretation of the answer (9 value closest to the patient on the left) is the following: when comparing the workload associated to the complete assistance of a priority 5 patient who has been discharged after a first consultation with the physician to that associated to the complete assistance of a priority 3 patient who needs further treatment, the respondent considers that the latter is extremely greater than the former.

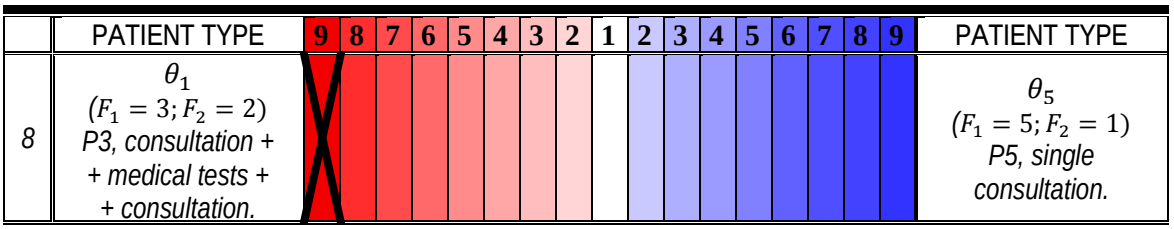


Figure 3.10. Example of pairwise comparison response in the questionnaire.

The complete questionnaire with all comparisons and the instructions are shown in Appendix E. Each questionnaire consists of all (in this case 15) pairwise comparisons as the one represented in Figure 3.9. That is, each expert is asked to compare the workload associated to two different type of patients identified.

Step 4. Selection of experts, dry run exercise, expert training, and elicitation session

The process related to the selection of experts, dry run exercise, expert training and elicitation session is the same as followed by the development of the stress function. It is detailed explained in Step 4 of Section 3.2.1.

The instruction sheet provided to experts and the complete questionnaire are shown in Appendix D and Appendix E respectively.

3.4.2 Phase 2. Data analysis

Step 5. Table of data. Consistency analysis for each expert's answers

In this section, the internal consistency of respondents are analysed according to Saaty's consistency ratio (CR) [119]. Workload scores coming from inconsistent physicians have to be reconsidered or revised, or if previous options are not possible, discarded.

First the pairwise comparison matrix, $A = [a_{ij}]_{n\times n}$, is normalized by equation (1) and then the vector of weights is computed on the basis of Satty's eigenvector procedure by equation (2)

$$a_{ij}^{*} = \frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}} \quad \forall\, j = 1, \dots n \tag{1}$$

$$w_i = \frac{\sum_{j=1}^{n} a_{ij}^{*}}{n} \tag{2}$$

Then, CR can be calculated using equation (3)

$$CR = \frac{CI}{RI} \tag{3}$$

where RI is the random consistency index obtained from a randomly generated pairwise comparison matrix and CI is the consistency index (CI) for each matrix of order n. Table 3.10 shows the value of the RI from matrices of order 1 to 10 as suggested by Satty (1980) and CI can be obtained from equation (4)

$$CI = \frac{\lambda_{max} - n}{n - 1} \tag{4}$$

where λ_{max} is an important validating parameter in AHP used as a reference index that represents the relationship showed by Satty (1980) between the vector weights, w, and the pairwise comparison matrix, A, as shown in equation (5)

$$Aw = \lambda_{max}\, w \tag{5}$$

Table 3.10 Random Inconsistency Indices (RI) FOR N = 10 ([119])

N	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
RI	0.00	0.00	0.58	0.9	1.12	1.24	1.32	1.41	1.46	1.49

Physicians whose CR are below a certain threshold are excluded. In the AHP method, if CR < 0.1, then the comparisons are considered to be acceptable.

Step 6. Aggregation of expert's answers

Once we have the pairwise comparison of all consistent individuals involved in the decision problem after the previous step's calculations, it is necessary to obtain an aggregate measure of them by the geometric mean of the individual assessments using equation (6).

$$a_{ij}^{g} = \sqrt[Q]{\prod_{q=1}^{Q} a_{ij}^{q}} \tag{6}$$

Where a_{ij}^{q} is an element of matrix A of an individual $q;\ q = 1, \dots, Q$, and a_{ij}^{g} is the geometric mean of all individuals a_{ij}^{q}. These geometric means make up the aggregate pairwise comparison matrix, $A^{g} = \left[a_{ij}^{g}\right]_{n\times n}$. The group CR can be calculated as in previous section's equation (3).

Step 7. Estimation of the workload function based on pairwise comparison

The vector of weights is computed on the basis of Satty's eigenvector procedure by equation (2) of step 4 after normalizing the aggregate pairwise comparison matrix, $A^{g} = \left[a_{ij}^{g}\right]_{n\times n}$ by equation (1).

This weights $w_1, \dots, w_6$ are the estimated values of the workload associated to the different type of patients $\theta_1, \dots, \theta_6$. Thus, the completed workload by a physician k from the beginning of the workshift at a time t is:

$$C_k(t) = C_k(p(t)) = \sum_{i=1}^{6} n_{ik}(t)\theta_i \qquad\qquad t \in \left(0,\ t_{END}\right]$$

where $n_{ik}(t), i = 1, \dots, 6$ represents the number of patients of each type i who have been treated and discharged by physician k from the beginning of the workshift until time t. These integer variables describe the completed workload scenario S.

3.5 Workload assessment in a shift: a case study

3.5.1 Phase 1. Preparation for data acquisition

Identifying the set of factors and their categories. Definition of completed workload scenarios. In the first step, we widely discussed with the ED physician staff in order to define every factor affecting the different workload associated to patient assisted, as well as all their possible combinations to pinpoint the workload variables. All experts – one is a member of our research group and has more than 10 years of experience working in the ED – agreed that the workload completed by a physician should consider the sum of all treated and discharged patients of each type.

We designed a common questionnaire containing the pairwise comparison of all types of patients which were essential to derive the normalized weights for the labour developed by a physician. From each expert of the panel, we were able to obtain the 15 comparison scores.

Drawing up the questionnaire for stress assessing. As mentioned in Section 3.4.1, the questionnaire contains one row for each pairwise comparison (see Figure 3.9), in this case 15. The two types of patients whose workloads are being compared are in both ends of the row. Between them there are a symmetric scale of 1 to 9 with a common origin that represents an equal workload associated to the assistance of both type of patients. Any other value represents dominance of one of them.

The complete questionnaire can be seen in Appendix E.

Selection of experts, dry run exercise, expert training, and elicitation session. This phase for the completed workload assessment was carried out in conjunction with that of the job stress (see Section 3.3.1). However, we got 42% of the ED physicians staff to answer the questionnaire, which were less responses than for the job stress questionnaire. Moreover, one of the experts only responded half of the questionnaire.

The final panel was made up of a total of 18 ED physicians: 6 with more than 15 years of experience, 5 with 5-10 years of experience, one with less than five years of experience, and the rest with unknown experience.

3.5.2 Phase 2. Data analysis

Consistency ratio of the raters. We first analysed each physician's response in order to detect inconsistent experts and as every response was anonymous and cannot be associated to their specific rater, their opinions were discarded. Saaty ([119]) considers a value below 0.1 to be acceptable, thus we discarded participants 5, 6, 13, 17, and 18 from the study (see Table 3.11).

Table 3.11. Saaty's consistency ratio (CR) of each workload questionnaire participant ([119])

	Participants																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CR	0.03	0.03	0.02	0.06	0.11	0.12	0.05	0.03	0.04	0.06	0.09	0.04	0.14	0.08	0.04	0.07	0.49	-

Aggregation of expert's answers. Once we had the pairwise comparison matrices, $A^{k\in K} = [a_{ij}]_{n\times n}$, of the most consistent individuals $1, \dots, K$ involved in the decision, we made up the aggregate pairwise comparison matrix, $A^g = [a_{ij}^g]_{n\times n}$ and calculated the group CR, which had a value below 0.02.

Estimation of the workload function based on pairwise comparison. The vector of weights $w_1, \dots, w_6$ was computed by equation (2) of step 4 after normalizing the aggregate pairwise comparison matrix, $A^g = [a_{ij}^g]_{n\times n}$. These were the estimated values of the workload associated to the different type of patients $\theta_1, \dots, \theta_6$ (see Table 3.12)

Table 3.12. Regression coefficients

Model variables (combining workload factors)		Normalized values
Care needs (F_2)		
Patients who only need a consultation ($F_2 = 1$)		
Priority 3 patients	θ_1	0.104
Priority 4 patients	θ_2	0.059
Priority 5 patients	θ_3	0.050
Patients who need more than one consultation ($F_2 = 2$)		
Priority 3 patients	θ_4	0.407
Priority 4 patients	θ_5	0.221
Priority 5 patients	θ_6	0.159

Thus, model representing the completed workload by a physician k from the beginning of the workshift at a time t is:

$$C_k(t) = C_k\big(p(t)\big) = \sum_{i=1}^{6} n_{ik}(t)\theta_i =$$

$$= 0.104n_{1k}(t) + 0.059n_{ik}(t) + 0.050n_{ik}(t) + 0.407n_{ik}(t) + 0.221n_{ik}(t) + 0.159n_{ik}(t)$$

$$n_{ik}(t) \in \mathbb{N}$$

$$t \in (0,\ t_{END}]$$

where $n_{ik}(t), i = 1, \dots, 6$ represents the number of patients of each type i who have been treated and discharged by physician k from the beginning of the workshift until time t. These integer variables describe the completed workload scenario S. This model allows us to assess the workload developed by every physician at any moment of the workshift through the completed workload information of the physicians' whiteboard (patients assigned and already discharged). It would also be useful to determine the differences in workload among physicians at the end of the shift.

Model validity was also checked by the ED physicians, who were asked to test the results and gave their approval.

Table 3.12 clearly shows the factor's influence on physician stress through the variable's coefficients. For the factor F_2 (care needs), patients who had been discharged after a first consultation ($F_2 = 1$, variables $\theta_1, -\theta_3$) are identified to be associated to less workload than those who needed a consultation that resulted in medical test requests, and needed a second consultation to be revaluated by the physician before being discharged ($F_2 = 2$, variables $\theta_4, -\theta_6$). Thus, $((\theta_1 - \theta_3) > (\theta_4 - \theta_6))$.

Within these two groups of patients factor priority (F_1) is in a descending order (inverse to severity index). The results –similar to job stress study's in Section 3.3– state that the most severe a patient is, the more workload he/she logically produces for the physicians: $(\theta_1 > \theta_2 > \theta_3)$ and $(\theta_4 > \theta_5 > \theta_6)$.

3.6 Discussion and conclusions

The medical literature recognizes that a better distribution of work among professionals reduces the level of stress and, therefore, mitigates the phenomenon of burn-out, which is so frequent in the health field and which results in a worsening of the health treatment received by patients. In fact there are several studies which support that high levels of workload and stress contribute to the high human and system error rates (e.g. [122]). For this reason, it is important to include indicators on the working conditions of physicians in the set of criteria that govern the management of patients. Therefore, the results presented in this chapter have an impact on the improvement of the physician's working conditions and the management of the patient flow.

In this chapter, we propose a new methodology in order to assess a physician's stress while working in the ED, taking into account workload, time pressure and uncertainty at any moment in a work-shift. That is, it objectively evaluates a situation through the workers' consensus. Contrary to any other stress measurement method, such as Dundee Stress State Questionnaire DSSQ, this is not subjected to a person's mood, age, sex, or other personal biases. For example,

Matthews [78], [123] has a general metric for evaluating the impact of environmental and personal factors on operator stress.

In this chapter, we have introduced a methodology that allows us to monitor the physician stress in real-time due to the workload and its characteristics spanning the work-shift. It takes into account not only the patients' priority and quantity but also the attendance phase, waiting time, etc. It also considers the importance the physicians have consensually given to the different stress factors for aggregating each patient's contribution to the job stress. Contrary to any other stress measurement method, such as Dundee Stress State Questionnaire DSSQ, this is not subject to a person's mood, age, sex, or other personal biases (see for example [78], [123]).

We incorporated to our concept of stress not only the workload but also the time pressure and uncertainty associated. For example, a priority 1 patient - assigned to a physician - contributes differently to stress in the following situations:

- Situation 1: waiting for C1 for 2 minutes.
- Situation 2: waiting for C1 for longer time than the time limit for priority 1.
- Situation 3: waiting for test results ordered by a physician in a previous C1 to be discharged.

In situations 1 and 2 there is uncertainty: the physician has not seen the priority 1 patient yet and does not know the medical care they require, the severity, or circumstances, etc. Thus, the workload associated to the physician in these situations entails a greater uncertainty -and consequently, it is more stressful- than in situation 3, in which a physician has already seen the patient and could have requested some medical test. Furthermore, situation 2 is even worse in terms of stress than situation 1 as patient's waiting time limit for the C1 has been exceeded, and their health status may have worsened or changed (time pressure).

The perception of all these nuances is possible because this method is based on the elicitation of experts' opinion and experience. A respondent of a self-report has conscious awareness of the experience of stress and can presumably report the feeling of this experience. This concept is called the perceived stress (Lindsay & Norman, 1980). Moreover, the subjective techniques are the least intrusive, the most flexible, the most convenient, the least time consuming, and the least expensive form to evaluate the stress.

Previous considerations allow us to assess the physician's job stress in every possible situation of the ED though the workload information of the physicians' whiteboard. The job stress score changes dynamically as the workload assigned to a physician changes during the work shift (existing patients health status evolves, new patients arrive, etc.). As an example, Figure 3.11 represents the dynamic instantaneous real job stress level experienced by the different physicians during their work shift (historical data from a Monday from 8:00 to 15:00 in the ED of the HCN).

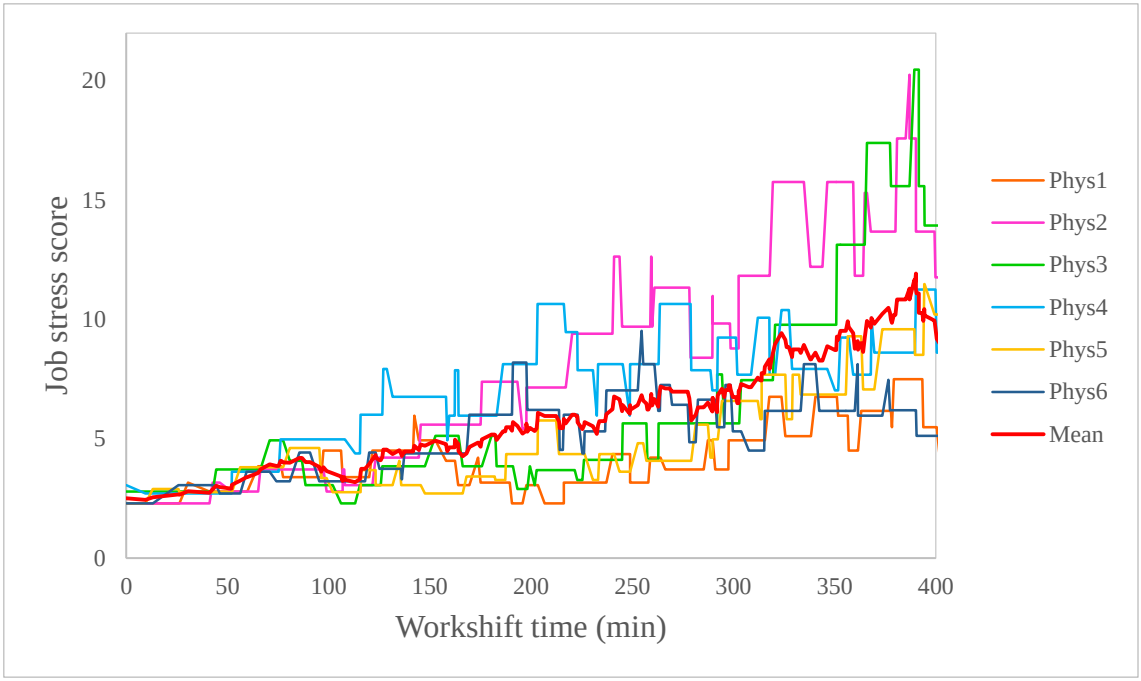


Figure 3.11. Stress associated to each physician along a specific work shift

It graphically confirms the feelings of stress and workload inequities among physicians reported by the HCN-ED physicians, which is neither healthy nor fair. There are situations in which a physician could have accumulated many patients as they were all very complex and required a lot of medical care, while other physicians are not stressed as they were only assigned very mild patients (see Figure 3.11). These results motivate the investigation to change the distribution of patients among physicians rule in order to reduce stress variability among physicians during the work shift and consequently, improve patients' quality of medical care and avoid the physicians' health problems. This problem is addressed in next Chapter 4. Moreover, the obtained job stress score – apart from the importance of being considered as a Key Performance Indicator (KPI) as in this case, it can also be used as a criteria to manage the ED patient flow.

All these remarks can also be extensible to the second case study of the methodology, in which the completed workload function provides a score of the effort made by a physician from the beginning of their workshift until any time t associated to the patients completely assisted and discharged by them. It allows us to assess the variability in the completed workload among physicians at the end of the shift due to the mix of workload assigned to them.

Different EDs can have a different mix of patients, different layouts, different staff policies, etc. that can affect how physicians experience the job stress produced by a workload scenario. Nevertheless, the methodology proposed in this chapter is general enough to be adapted to monitor the physician's job stress of any ED. It only needs physicians to assess different workload scenarios following steps 1-4 (these scenarios will be created by the design of

experiments taking into account the stress factors identified as relevant by the ED physicians) and then the stress function is estimated by analyzing the data following steps 5-7. Therefore the presented methodology, with the help of the case studies presented here and the supplementary material (questionnaires and guidelines included in Appendix A and Appendix B, and mathematical details included in Appendix C) can be applied to monitor physicians stress in any ED. Moreover, EDs with similar ways of operating, type of patients, etc. to those explained in this study can directly apply the obtained regression function to monitor their physician's job stress.

It is important to emphasize the importance of the qualitative validation of the obtained stress function by the ED physician as it will be used in the workload assignment. In the case study, the obtained results from the statistical analysis for the stressors importance were in general what physicians expected. Patients who have not yet been seen by a physician for the first time contribute to higher levels of stress than the others and within that group, those whose waiting time has been exceeded contribute the most. It is remarkable that during Phase 1, all ED physicians reported that resident supervision was an important stressor but the results did not show statistical significance. Nevertheless, in the validation phase, we all agreed that it must be included in the regression function as they consider the measure to be fairer to be used in the ED.

Chapter 4 Patient flow management from triage to treatment.

4.1 Introduction and related literature

EDs are widely known for their stochastic nature, unpredictable arrivals and—in the last two decades— serious, growing overcrowding problems [124]–[126]. The net effects of ED crowding include poor patient outcomes [127]–[130] long waits to be seen [128], [129], patient dissatisfaction and patient complaints [124], [128], [131], ambulance diversion, [128], [130], [132] and increased number of patients who leave without being seen [130], [132]. Moreover, studies have identified the LOS in the ED as the most important determinate of patient satisfaction [133], [134], which often declines when waiting times increase too [135]–[139]. Changes to front-end operations are of particular interest to reduce ED crowding because they are usually beyond the direct control of the ED (physicians, nurses, and administration) without requiring the involvement of external stakeholders, which would be practically or politically more difficult.

Moreover, the fluctuant nature of the ED is sometimes coupled with punctuations of high-risk time-critical activities, which could not only put patients' safety at risk but also cause stress to physicians. ED Physicians are usually exposed to more severe stress than other departments' physicians [77] whose principal sources are time pressure, critical decisions and amount of work [77]. Thus, it is important to highlight some of the forefront issues that the emergency medicine community should address which are the overcrowding, efficiency, and patient and provider safety to avoid serious consequences.

As explained in the introduction of this part of the thesis, patients access to the examination room, where a triage process classifies the patients according to their severity. Triage may include performance goals in terms of the percentage of patients who should have access to the physician consultation before certain time limits. These time limits and percentages will vary according to each type of triage-level (see Table 2.2). After triage, all patients wait in a queue for consultation and evaluation by a physician. This physician will be responsible for servicing all the needs of the patient assisted (clinical tests, such as blood tests, X-ray, scan, specialist's

consultation, etc.) until they are discharged from the ED (to a hospital ward or to the patient's home).

In this chapter the first phase of sequencing decisions is addressed, which is the determination of the order/priority in which patients are initially taken from the waiting area to start treatment with a specific physician.

Many departments use the traditional physician self-assignment process, whereby physicians assign themselves to (or "pick up") patients at their discretion. Physicians may base their decision to pick up a new patient on their perceived capacity to treat another patient, how ill the patient is, the perceived needs of the department (with respect to whether it is busy or not), among other factors. In other EDs (see for example [60]), triage nurses are the decision makers who establish who is the next new patient assisted by a physician, and physicians decide when they have enough capacity to assume the new patient's treatment. Newly registered walk-in patients are placed in a queue of ready-to-be-seen cases by triage nurses, which allows physicians to "sign up" for patients when they feel ready to see their next patient. Both these systems are considered "pooling system" in which there are a shared queue that leads to multiple servers (physicians).

There is also another possible configuration in which there are one dedicated queue that leads to each server. By using this configuration, incoming patients are immediately assigned to a physician after triage by a rotational rule (see for example [140]). Once this assignment occurs, the physician can see the assigned patient –even if he/she has not been seen yet - listed under his/her electronic tracking board when logged onto the patient management system.

On the one hand, ED crowding may be improved by reducing waiting times for the placement of patients in ED beds. By doing this, patient care along with prompt discharge will initiate earlier [131]. These also would increase patient's quality of care and satisfaction. On the other hand, the randomness of patient arrivals may result in inequality of stress experienced by physicians due to their associated workload. This relevant problem may influence the service quality and physician's working conditions, and should also be considered.

The main purpose of this chapter is to analyze different ways of managing patient flow, and to design and implement a new automatic and real-time tool to assist the triage nurses in the decision making process. Testing of new management policies in complex systems, where the results are unpredictable, especially in those systems where there is a strong ethical component such as the health system, is carried out by means of simulation techniques [42], [141]. The two objectives are: optimizing waiting time and quality of care (patient's perspective) and optimizing physician's workload balancing and working conditions (physician's perspective). This is the first time that the latter objective is considered in ED "front-end" operations. Moreover, the study will also be applied to a real setting, the ED of the Hospital Complex of Navarre (HCN).

4.2 The patient-to-physician allocation problem

4.2.1 One queue vs multiple queues

The two common queue configurations mentioned above to manage patient flow determining the order as well as assigning physician that will treat them are: 1) multi-server single-queue (SQ) in which there is a shared queue with all patients that leads to multiple physicians, and 2) multi-server parallel-queue (PQ) in which there is one dedicated queue that leads to each physician.

Traditional queueing theory demonstrated through analytical models that a pooled queue configuration (SQ) is more efficient than dedicated queues configuration (PQ) [142]–[144]. By allowing patients to be served by any available physician rather than having them wait for a specific server to become available, pooled queue configurations help mitigate the negative effects of variability in arrivals, which leads to shorter waiting times for service, less expected throughput times, and less expected work-in-process (WIP) [144]–[148]).

However, recent empirical works of medical literature suggest that the previous statement may not always be true in practice. Longer average waiting times and lengths of stay are experienced by patients when physicians are assigned patients under a pooled queueing system as opposed to a dedicated queueing system of rotational patient assignment ([135], [136], [140], [149]–[151]). There are four groups that have previously reported some version of a system in which patients were assigned to alternating teams (PQ) [135], [136], [149], [150]. Three of these groups reported a decrease in arrival to provider time (of 9.5 minutes [149], 13 minutes [136], and 4 minutes [150]), two reported an increase in patient satisfaction [135], [136], one reported changes in length of stay (of 39 minutes) and discharge rate (of 1.05, 1.07, and 1.05 greater during the second two hours, in the penultimate two hours, and in the final two hours of a physician's main ED shift respectively) [150]. An older report describes a semicontrolled study in which rotational patient assignment was applied for residents on the "medical side" of a "medical side/surgical side" ED at a teaching facility [151]. LOS improved on the medical side by approximately 15%, whereas LOS increased on the surgical side [135]. And finally there is a study in which patient are assigned to physicians rotationally reporting an improvement in LOS, APT and LBBS and complaint ratio.

According to recent studies, this happens because queue configuration may make an impact on the physicians' behaviour ([150], [152]–[157]), which is not considered in most widely used queueing models in both academic literature and practice. These models assume that the service rate is exogenous, which, while credible for nonhuman servers, is problematic for human servers (see for example in other contexts [158], [159]). Do et al. [160] and Armony et al. [161] start to incorporate these behavioral aspects.

In a SQ system in which physicians assign themselves to patients at their discretion, they perceive their patient loads differently and "pick up" additional patients at varying rates [135]. These rates are related to varying factors such as each individual physician's tolerance of workloads, departmental and management expectations, monetary remunerations, etc. Usually, when there are many patients waiting to be seen, quicker and more efficient physicians see more patients, which is perceived as unfair and may negatively impact on the speed at which servers work [162]. This occurs in many public hospitals, where physicians are salaried.

Shunko et al. [152] rely on behavioral experiments to show that workers process items at a slower rate in SQ systems than in PQ systems. The slowdown in processing in the SQ system is directly attributed to the effect of social loafing in a shared workload environment. As it is also demonstrated by Wang and Zhou [153], the server may slow down. This happens when the workers have the opportunity to free ride on others' by reducing their share of the work and the effort required to perform this work at the expense of their colleagues. Under this configuration (SQ), waiting patients remain no one's direct responsibility until an available physician takes ownership of them ([135], [136]) resulting in delays in care since there is little or no reason for a physician to electronically claim a patient in the waiting room.

On the contrary, in a PQ system, the responsibility for each patient care in the ED waiting room is assigned to a specific physician, who have more ownership of patients, thus, more actively manage the patient flow. This is an incentive for them to initiate evaluation more promptly and make every effort to get patients seen within their time limit ([155], [162]). Moreover, enabling performance comparisons across teams or physicians with the full visibility into each physician's length of the queue, whose workload is tracked on the ED electronic board, also contributes to making the physicians more efficient (see for example [163]). In a series of behavioral experiments, Shunko et al. [152] also show that servers work faster when performance feedback is made more salient by increasing the visibility into the length of the queue.

Armony et al. [161] attribute this phenomenon to the servers' degree of discretion over their choice of service capacity and their type of work aversion (averse to high levels of workload, and/or preference for idleness over occupation), which is also the case of a SQ in the ED.

Additionally to ED performance superiority, from a medical point of view, in a SQ configuration there is little clarity of the physician's responsibility for ready-to-be-seen but unassigned waiting patients. On the contrary, in a PQ system, the physician's ownership is nearly immediate and always unambiguous. This allows triage nurses to identify a responsible physician when one is needed to guarantee patient's safety in case they get worse before the first consultation. Moreover, physicians know which waiting room patients they are responsible for enabling them to initiate earlier "preorders" on waiting patients before consultation.

4.2.2 Key Performance Indicators (KPIs)

Patient's perspective: safety and quality of service

Welch et al. [72], Welch et al. [73], and most recently Vanbrabant et al. [14] list various metrics by which ED performance can be measured, such as the arrival to provider time (APT, or "door-to-doc" time). This important time interval is widely used in emergency healthcare services, since many illnesses are time-dependent, and a delay in the diagnostic evaluation by a qualified medical provider could be a health risk for the patient. Most EDs define a maximum waiting time for each acuity level and set performance goals related to them, as explained in Table 2.2's CTAS. The ratio of patients whose APT exceeds the time limit is also considered a KPI.

There are also other important time related measures influenced by the patient flow management policies, such as the arrival to discharge time, called the "length of stay" (LoS), which has an impact on the patient's quality perception of the received healthcare service. In this study, from the patient's point of view, mean APT, mean ratio of patient exceeding the time limit, and mean LoS for patients of each priority will be considered KPIs.

Physician's perspective: working conditions

According to recent research, the time it takes a resource to care for a patient is not independent of the state of the process including the current workload [153]. Thus, they make dynamic service capacity adjustments in response to varying levels of workload [164]. This can also be indirectly affected by queue configuration, which influences the queue length that the server faces. The servers' responses to increased workload could vary [164] and affect quality of care as well as physicians stress.

Across a variety of service settings, prior work has shown that varying levels of workload may lead to increasing ([165]), decreasing ([166], [167]), inverted U-shaped ([159], [168], [169]), or N-shaped ([170]) responses of service time. Other studies show that quality may suffer due to load ([168], [169], [171]), or that workers may burn out due to load ([172]). A key assumption in this line of work is that as individuals experience more load, they choose to work faster in the short-term, although this speeding up may negatively impact performance in the long-term.

As explained in Chapter 3, the instantaneous workload is all patients a physician is managing simultaneously and as we also demonstrate, high patient loads increase physician stress [84], [173], and high priority patients too.

In this study, from the physician's point of view, the average of stress per physician during the work shift, $\bar{Y}$, will be one of the KPIs as well as the variability of stress among physicians during the work shift by using the mean square error (MSE) along the shift $\overline{MSE\ (Y)}$:

$$MSE\left(Y(t)\right) = \frac{1}{Q}\sum_{k=1}^{Q}\left(\hat{Y}(t) - Y_k(t)\right)^2 \qquad t \in [0, t_{END}]$$

$$\overline{MSE\ (Y)} = \frac{1}{t_{END}}\int_0^{t_{END}} MSE(Y(t))dt$$

where Q is the number of physicians working simultaneously in the ED , $Y_k(t)$ is the stress experiencing by physician k at time t, $\hat{Y}(t)$ is the estimated average stress for physician at time t, and t_{END} is the duration of the shift. The stress is measured with the method developed in Chapter 3.

The variability in the number of patients of each priority assigned to each physician along the workshift will be also used as a KPI by using again the MSE:

$$MSE\left(N_i(t)\right) = \frac{1}{Q}\sum_{k=1}^{Q}\left(\hat{N}_i(t) - N_{ik}(t)\right)^2 \qquad t \in [0, t_{END}]; i = 1, \dots 3$$

$$\overline{MSE\ (N_i)} = \frac{1}{t_{END}}\int_0^{t_{END}} MSE(N(t))dt \qquad i = 1, \dots 3$$

Where $N_{ik}(t)$ is the number of patients of priority i actually assigned to physician k at time t, $\hat{N}_i(t)$ is the estimated average number of patients of priority i per physician at time t, and t_{END} is the duration of the shift.

4.2.3 Definition of rules for patient-to-physician allocation

As previously mentioned, the medical literature recognizes that a good distribution of work among professionals reduces the level of stress and, therefore, mitigates the phenomenon of burn-out, which is so frequent in the health field and which results in a worsening of the health treatment received by patients. In fact there are several studies which support that high levels of workload and stress contribute to the high human and system error rates (e.g. [122]).

For this reason and for countering randomness in arrival and fluctuant nature of the ED, it is important to design patient flow management rules and to include indicators on the working conditions of physicians not only in the performance assessment but also in the set of criteria that govern the management of patients.

In this section, we propose several patient-to-physician allocation rules (PPAR) based on different criteria to be analyses. Advanced algorithms are commonly used to solve similar problems in resource allocation in other industries, but not in healthcare. For example fair, efficient, and skilled-based routing incoming calls in call centers are well-studied mechanisms that determine the optimal assignment of incoming calls to agents [148], and data

communication networks and manufacturing systems use algorithmic scheduling techniques to ensure efficiency of flow [174].

Single Rotational Rule (SRR)

Currently, there are some EDs - including the ED where the studies are carried out explained in Chapter 2– that has a single rotational patient assignment system, that is, patients are assigned to a specific physician rotationally as they are triaged after arriving to the hospital ED. Some studies have demonstrated that this SRR assignment results in ED performance superiority to the single queue system without previous assignment (see [135], [136], [140], [149]–[151]). Moreover, it facilitates the triage nurses' decision making and ensures an equal number of assigned patients to different physicians.

However, the differences in the complexity of clinical cases lead to an unbalanced physicians' workload since the average severity of patients assigned to one physician might be higher than that of patients assigned to another ([140], [151]). Thus, to overcome this handicap, we have defined other different patient-physician assignment rules.

Multiple Rotational Rule (MRR)

The MRR is an improved extension of the SRR that takes into account patient's ESI (1-5) attributed by the triage nurse. High and low severity patients are considered and the rotational assignment rule is applied to each of these patients' categories.

Figure 4.1 shows the different physicians' queues (patients associated to them) obtained by applying the SRR and the MRR to the same 18 incoming patients sequence. These patients are on the top of Figure 4.1 and they have their priority assigned by the triage nurse.

On the lower left-hand of Figure 4.1 a SRR from physician 1 to 5 has been applied to assign patients to physicians. Even if all physicians have almost the same total number of patients to assist (3 or 4), the randomness of incoming patients causes an unbalanced workload. For example Physician 2 has 4 low priority patients assigned while Physician 1 has 3 high priority patients assigned.

On the lower right-hand of Figure 4.1 a MRR has been applied to assign patients to physicians. High priority patients have been assigned rotationally from physician 1 to 5 while low priority patients have been assigned rotationally from physician 5 to 1 (opposite sense). In this case, the workload assigned is better balanced as all the physicians have an equal number of patients of each priority assigned.

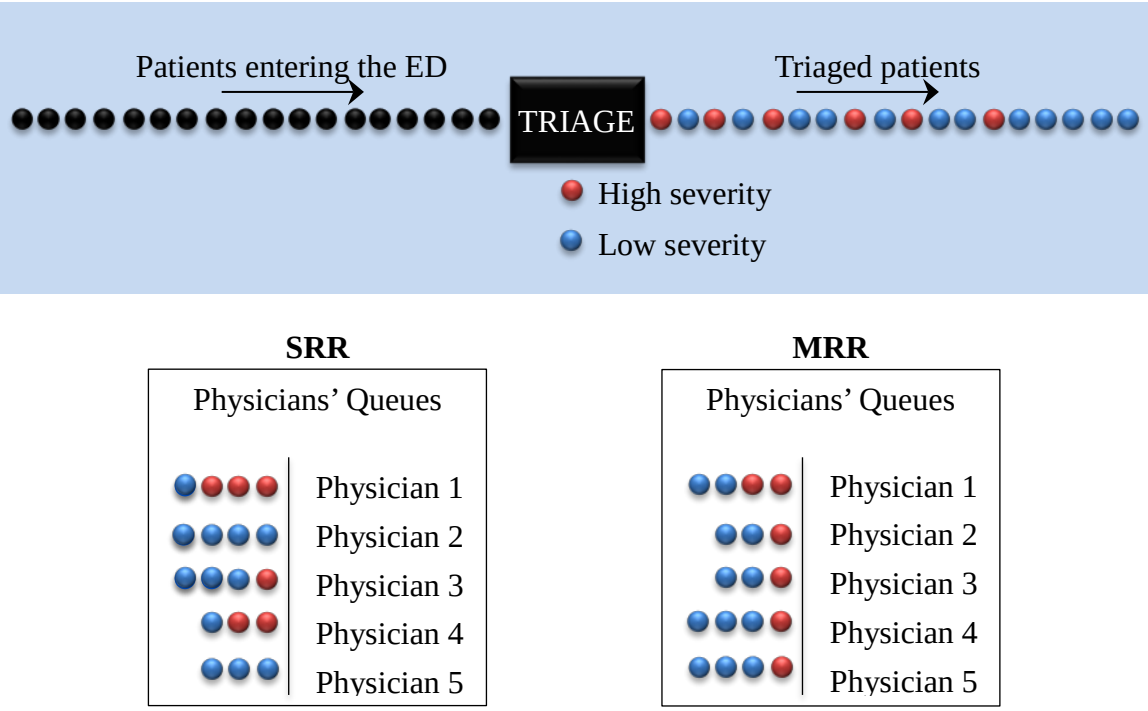


Figure 4.1. Resultant physicians' queues by applying two different assignment rules to the same patients.

However, the MRR assignment rule can also cause significant differences among physicians' job stress as this assignment rule does not consider complexity of patients or physicians' remaining work burden at the time of a new patient allocation. There are situations in which a physician accumulates many pending patients as all the assigned patients are very complex and require a lot of medical care, while other physicians are idle as they are only assigned very mild patients. To overcome this problem and reduce inequities among the physicians, we propose the next patient-to-physician allocation rule.

Physician's stress balancing rule (SBR)

The physician stress balancing rule (SBR) takes into account all the information of physicians' pending patients at the moment of a new patient's allocation in order to reduce job stress variability among the physicians during the work shift.

This rule uses as criterion the physician's work conditions' indicator developed in Chapter 4 and assign each patient upon arrival to the physician i with the lowest job stress score verifying

$$i = arg\ \min_k\{Y_k(t)\} \quad \forall \quad t \in (0,\ t_{END}],\ k \in [1, Q]$$

where $Y_k(t)$ is the stress being experienced by physician k at time t. The job stress experienced by a physician is measured with the method developed in Chapter 4 as a function of the number of pending patients of each priority assigned to the physician, their stage in the medical care

process (uncertainty associated), waiting time targets (time pressure associated), and teaching duties responsibilities. That is, the scenario represented by the vector $X_1, \dots, X_{11}$.

Physicians' stress and completed work balancing

Previous PPAR, the SBR, assigns patients to physicians as they arrive to the ED by taking into consideration pending patients at the moment of the allocation. This rule has been criticized by physicians because it can cause inequities of the total workload assigned to each physician at the end of the workshift due to the fact that not all physicians work at the same rate. Moreover, as mentioned in Section 4.2.1, servers' behaviors can be influenced by the queue structure, that may slow their service rate when workers have the opportunity to work less at the expense of their colleagues' effort. By using the SBR, the less patients a physician discharge, the more patients they accumulate, and the less probability of being assigned a new patient.

To counter this possible behavior, we propose a new rule based not only on current pending patients but also on already discharged patients. Thereby, it also considers the labor developed by each physician from the beginning of the workshift until the moment a new patient arrives.

The physicians' pending workload stress and completed workload balancing rule (SWBR) is a parametric assignment rule obtained from the linear combination, $SW_k(t)$, of the two standards: job stress, $Y_k(\mathbf{t})$, and completed workload, $C_k(t)$ described in Section 3.4. An incoming patient is assigned to the physician i verifying

$$i = arg\ \min_k\{SW_k(t)\} \quad \forall \quad t \in \left(0,\ t_{END}\right],\ k \in [1, Q]$$

where

$$SW_k(t) = \lambda \frac{Y_k(\mathrm{t}) - Y_l(\mathrm{t})}{Y_j(\mathrm{t}) - Y_l(\mathrm{t})} + (1-\lambda)\frac{C_k(t) - C_g(t)}{C_h(t) - C_g(t)} \quad \forall \quad t \in \left(0,\ t_{END}\right],\ k \in [1, Q]$$

where $l = arg\ \min_k\{Y_k(t)\}$, $j = arg\ \max_k\{Y_k(t)\}$, $g = arg\ \min_k\{C_k(t)\}$, and $h = arg\ \max_k\{C_k(t)\}$.

4.3 Analysis of a multiple rotational rule in a real setting

In this section we analyze how one of the rules proposed in the previous section impacts on a real setting: the multiple rotational rule. The real setting is the ED of the HCN and it has been chosen because of the potential good behavior shown in the simulation model described in Chapter 2 that improves the current ED performance. This way, the real impact of the new management policy is demonstrated on the real ED, which proves the research carried out with

simulation models. It validates the mathematical models, the model completeness, and the methodology approach.

Similarly, this procedure may be reproduced for the rest of the proposed rules in Section 4.2.3 based directly on physicians' stress. In our case, this was not possible because we were unable to access the patient's data in real time. Next, each step of the methodology to implement the results is presented, followed by the description of the real implementation in the HCN.

4.3.1 Patient-to-Physician allocation problem at the Hospital Complex of Navarre (HCN)

During the day, as many other EDs, the HCN - which is described in more detail in Chapter 2 - organizes the patient care into two different care circuits: one for the more critical patients, i.e., circuit B (CB), and another for less critical patients, i.e., circuit A, (CA). They both have dedicated physicians, nurses and ancillaries that are not shared with the other patient care circuit. All priority 1 (P1) and priority 2 (P2) patients are assigned to CB while all priority 4 (P4) and priority 5 (P5) patients are assigned to CA. Priority 3 (P3) patients can be assigned to both depending on the illness. Triage nurses assigned patients to one of the two different care circuits and within the selected circuit they manually assign patients to a specific physician. The assignment is placed on the electronic tracking board, which is visually available to the entire department. Each physician has their own rack of patients and evaluates them at their own pace, with the understanding that they must see and evaluate all patients assigned to them until one hour before the end of the workshift. They try to avoid the physician that enters the ED the next workshift to take over their patients. The assignment of patients to a specific physician follows a rotational rule without considering their complexity, their priority, or the physician's pending patients. For example, 1st patient assigned to CA of the day (a P3 patient) is assigned to physician 1, 2nd patient assigned to CA of the day (a P3 patient) is assigned to physician 2, 3rd patient assigned to CA of the day (a P4 patient) is assigned to physician 3, and so on until the last scheduled physician in CA during the shift, and then the round starts again with physician 1. For patients assigned to CB the procedure is the same.

This rule facilitates triage nurses' work and guarantees an equal number of patients across physicians as their compensation model is salaried, with no component for clinical productivity. This rule has been proved to be the best until now in medical literature's interventions, contrary to theory that suggests a single queue configuration is better.

However, physicians in the ED of the HCN reported feelings of stress and workload inequities among physicians. This was confirmed in Section 3.6 by dynamically tracking the stress experienced in the workplace by each physician considering their workload information (patients assigned) during their work shift. This was done after developing the job stress score in real time by consensus. The results motivated the investigation to change the patient

assignment rule in order to reduce job stress variability among physicians during the work shift and consequently improve patients' quality of care.

In this study, we consider the management of the healthcare CA, of lower priority patients (levels 3, 4, and 5) as it was described to be the most crowded. Moreover, patients treated in care circuit B are too critical to take part in an experiment, who will be used as a control. CA has five exploration rooms and a senior physician in each exploration room. Patients who are treated in this care circuit are rotationally assigned from physician 1 to 5 just after triage (see Figure 4.2), as mentioned above. In the next section we outline the research carried out to improve the distribution of patients among physicians, from idea generation to implementation. This aims to optimize not only patient waiting time but also working conditions.

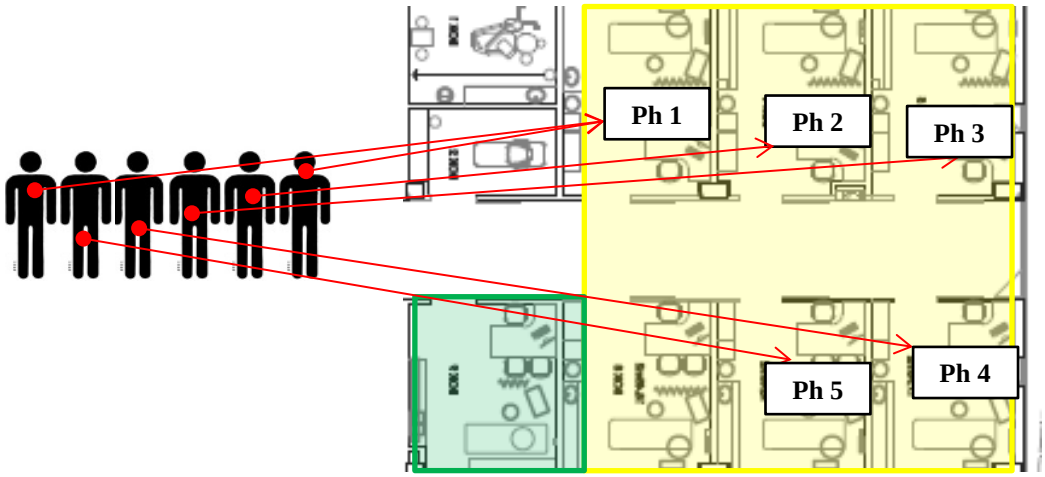


Figure 4.2. HCN CA assignment system.

4.3.2 Phases

The methodology followed in the improvement phase is structured in 9 phases and is summarized in Figure 4.3.

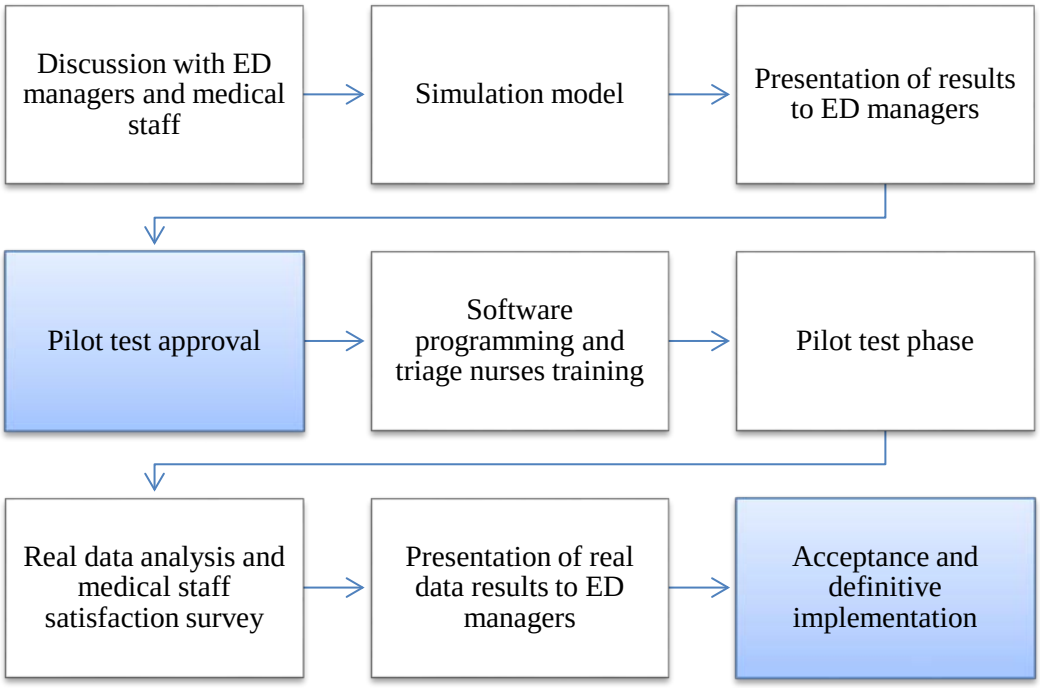


Figure 4.3. Summary of the methodology structure of the improvement process in the HCN

Discussion with ED managers and medical staff

First, a meeting with the ED managers is needed to agree that a solution to the problem expressed in previous Section 4.3.1 is needed. Then, it is important to organize several workshops with medical staff who directly face its consequences. It helps to collect as much information as possible about the processes, way they work, feelings, insights etc. The workers stated that the most urgent care circuit was the less critical patients circuit as it was more crowded. From their medical staff point of view, there are two very different types of patients in the less critical patients circuit. These are called high severity patients (priority 3) and low severity patients (4 and 5) and have very different provider time limit, need different resources, etc., which means, a different workload. Generally, physicians sort their patients first by acuity and then by time of arrival in order to see them for their first time within the time limit defined by the triage level.

Particularly in our case, they showed concern for the unfair distribution of patients among physicians displayed on the electronic tracking board during one of our workshops at the hospital. Figure 4.4 shows a screenshot of the real electronic tracking board. It was taken at 12:04, the day 06/05/2016 and the patients who had come to the system from 11:20 to 12:04 are listed. There is no personal data to be identified of the patients except age, visit motive, priority, care circuit associated and specific physician within the care circuit.

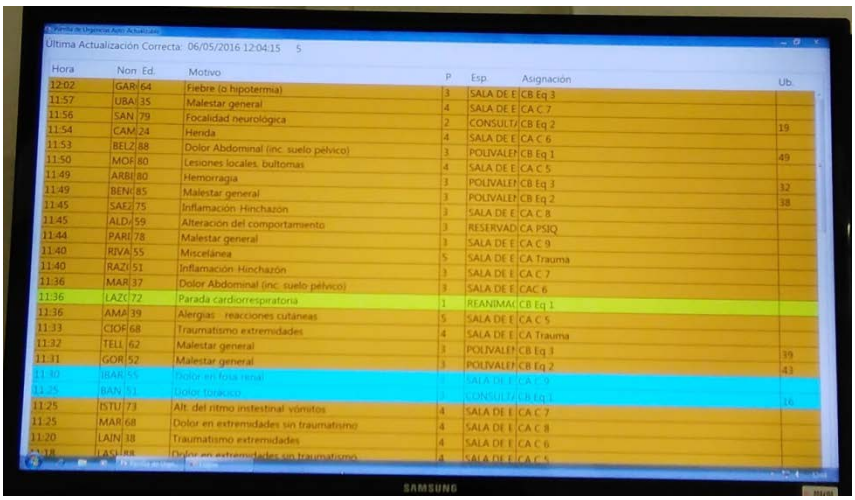


Figure 4.4. Screenshot of the real electronic tracking board of the ED, day 06/05/2016 at 12:04.

Table 4.1 summarizes the distribution of patients during the 42 minutes period represented in the screenshot of Figure 4.4. They discussed the mentioned situation that their colleges were having during the workshop. Physician 5 had 3 high priority patients assigned while Physician 1 had 0, who are considered to have much more workload associated to their care process by all of them.

Table 4.1. Distribution of patients during the period represented in the screenshot of Figure 4.4.

N. Of patients	Physician 1	Physician 2	Physician 3	Physician 4	Physician 5
Priority 3	0	1	1	1	2
Priority 4	2	2	2	2	0
Priority 5	1	0	0	0	0

During our workshops, we proposed some new rules to them (described in Section 4.2), which were discussed. They expressed the impossibility of accessing to the ED information system in real time at that moment in time, which is necessary to assess the job stress and calculate all pending and discharged patients. These calculations are required for the criterion of some of the proposed rules, which had to be discarded. Finally, we all come up with an easy, feasible, and reasonable policy that could possibly be implemented, which was the multiple rotational. In this case in which there are only two different types of patients, it will be named parallel rotational.

It is essential to keep medical staff motivated as participants during the whole project because they will be the actual receptors of the proposed improvement solution.

Simulation model

Once we have this new proposal for managing patients in the ED, the necessary methodologies and indicators have to be developed. In this study, it was essential to define and validate a job

stress score by the consensus of the physician of the ED to assess inequities in stress experienced by physicians associated to their workload in real time, which is describe in Chapter 3.

Then, a simulation model of the ED of the HCN (explained in detail in Chapter 2) was developed and validated with medical staff to theoretically test some new rules proposed in Section 4.2. In our case, in the previous phase they stated that only the multiple rotational rule could be implemented. This policy was tested in the simulation model, which proved it is significantly superior to the rotational rule that was used up until that time. The main results were a reduction in high priority APT (10%) while worsening APT of mild patients (6%), see Figure 4.5. Moreover the differences in work completed at the end of the shift was reduced by 6% (standard deviation).

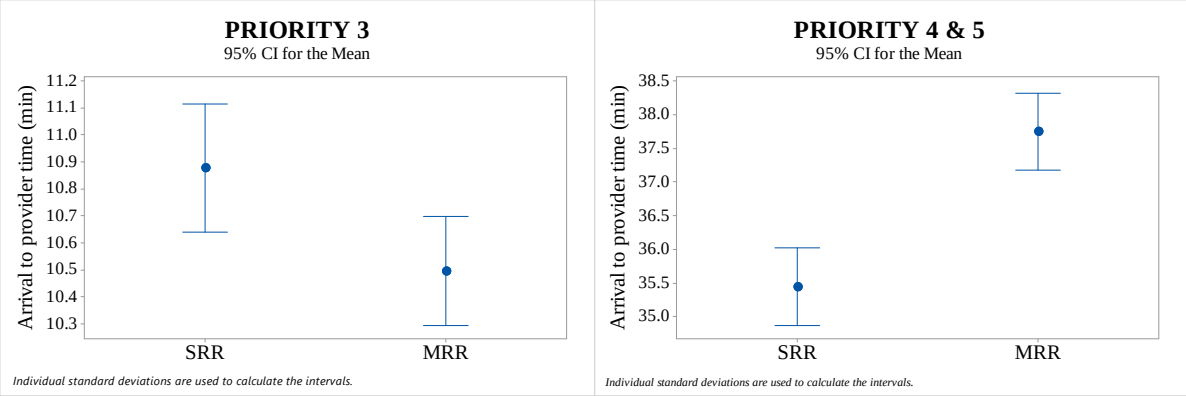


Figure 4.5. APT of high and low severity patients comparison single rotational rule, currently used, multiple rotational rule, proposal)

Presentation of results to ED managers and pilot test approval

The analysis of the new management policies must be presented to the ED managers as well as the medical staff who work daily in the department, who should agree with some of the new proposals and most importantly accept to perform a pilot test.

In our case, the results which are described in the previous phase were presented to the ED managers and medical staff, who approved a pilot test performance.

Software programming and triage nurses training

Once the pilot test is approved in order to assess one of the new assignment rules proposed in the real system, an easy to use software is necessary to be programmed in order to implement the selected solution. The final users should collaborate in its development to add some necessary features and suggest improvements. Then, it is essential to provide training sessions to all people participating to get them familiar before the implementation (see Figure 4.6).

Figure 4.6. Training session and user's manual and frequently asked question provided to medical staff.

In our case, the selected programming language was Java in order to make it visually pleasant and for free. During the process, the final users - the triage nurses - collaborated in the development to suggest some needed features. For example, the manual assignment of a physician in case they would consider it appropriate is also included (see Figure 4.8). In these cases, the algorithm has a memory and that will be taken into account when assigning the next incoming patients. There are other features like priority modification (see Figure 4.7, part 3) of a patient if they get worse where the memory is optional, or physician modification. Patients that abandon or change care circuit can also be considered to autoadjust the incoming patients' distribution.

As the software cannot access to the ED system in real time, at the beginning of the workshift the number of physicians available must be defined. Every time a patient arrives, it is necessary to insert patient code and priority (part 2 of the window shown in Figure 4.7) and the window shown in Figure 4.8 pops up to suggest the physician to whom the patients should be assigned.

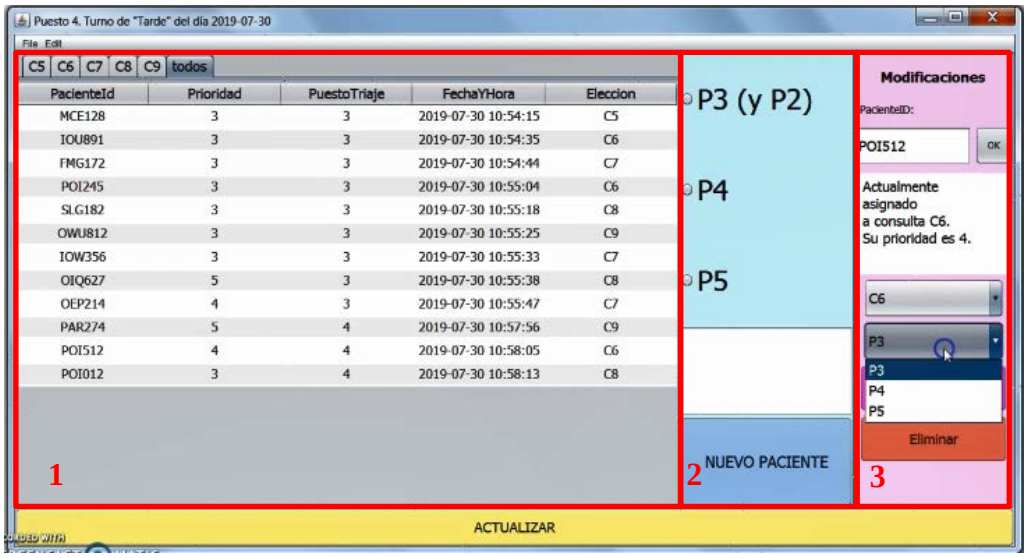


Figure 4.7 Main screen of the software. The panel on the left shows every patient and their assignment in the ED and per physician. The blue part in the middle is for introducing a new patient as they arrive.

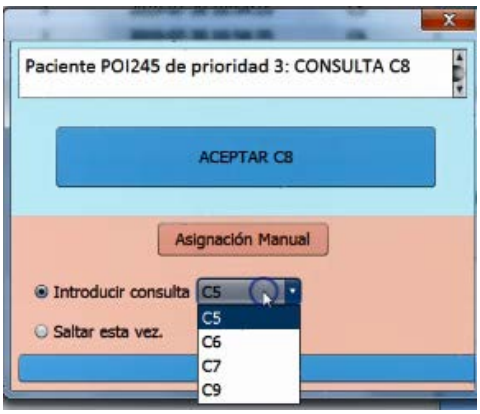


Figure 4.8 Screenshot of the suggested assignment by the software. In this case, the triage nurse can accept physician 8 (it starts counting physician from 5 on) or insert manually that it is more appropriate to send the patient to physician 5.

Pilot phase

During the pilot phase, medical staff, and particularly triage nurses use the new management policy implemented in the new software. It is essential that all medical staff are informed about the situation to let them know that physicians will receive patients in a different order than that they are used to.

It is also very important to provide support as well as keep in touch if any problem arises.

The pilot test in the case study ED was performed between 4th June and 2nd July (for a month). During this period, the software had the parallel rotational assignment rule implemented for patients assigned to CA. That is, there are two queues with all the physicians who are ordered opposite ways in each of them. Severe patients are assigned to one physician of the first queue on a rotating basis (for example starting from physician 1 to physician 5 and repeating this wheel along the day) while mild patients are parallel assigned to one physician of the second queue on a rotating basis (in this case starting from physician 5 to physician 1 and repeating this wheel throughout the day). As exposed in the previous phase, medicine is not an exact science, and medical staff want to control exceptions in which one patient should be specifically treated by one of the physicians. In this event, it is possible to manually assign a patient to that specific physician independently of what the software suggests, and then the software redress physicians workload with subsequent patients. Finally, as mentioned, within CB, the assignment rule did not change from the previous period.

Real data analysis and medical staff satisfaction survey

After the pilot test takes place, it is necessary to request administrative records of patients (the period before and after the implementation of the software) to conduct a before-and-after study in order to compare the previous and the new management rule. All the details of our study data analysis are in next Section 4.3.3 and the results were very positive.

A survey on medical staff satisfaction is also very important in order to collect their perceptions, improvements suggestions, problems, acceptance, and their feelings to subjectively assess the software and the management rule. This survey should be approved by the chief of the department for quality and development.

In this case study, the survey was approved by the nursing chief (as the final users were the triage nurses) and was administered to all staff that had worked at triage during the pilot test period. As an example, the survey developed is in Appendix G, in which several aspects are requested such as technical and visual aspects, consideration of permanent implementation of the new rule, and its extending to the other care circuit that did not participate in the pilot test.

Presentation of real data results to ED managers

Finally the results of the real ED system must be presented to the ED managers and the medical staff to objectively demonstrate the benefits of the new assignment rule to them. They have the option of accepting the permanent implementation of the software and its connection with their ED system.

4.3.3 Analysis of results

We report the results of transitioning from single rotational patient assignment to parallel rotational patient assignment taking into account patients priority at a single facility, the ED of the HCN, with the goal of reporting the operational metrics of LOS, APT, ratio of patients exceeding the APT limit, rate of early (within 72 hour) returns, and ratio of patients of each type assigned to the different physicians. We report these metrics while noting and accounting for several potential confounding variables.

Methods

Study design and setting

This is a retrospective before-and-after observational study in which we analyzed routinely gathered ED operational data in order to compare the effects of a parallel rotational patient assignment system.

As mentioned, in this study the new patient assignment rule is applied in the CA, while care CB performs as a control group. The board certified physician labor pool was constant throughout the study period, all of them worked during the entirety of both periods (simple rotational and parallel rotational patient assignment period) and had more than 3 years or more of postresidency emergency medicine experience. Each year at the end of June, new residents start the training program.

Throughout the study period, the ED used both the same electronic medical record and the electronic tracking board software. Moreover there were no changes in organization processes or triage system.

Selection of Participants

There are 28 days of parallel rotational patient assignment (June 4 to July 1) with 8892 patient visits. To minimize confounding of our current data the results are compared to the corresponding 28 days of previous years as seasonality is very important and residents gain knowledge through the course that starts in June.

We identified a matched day in the last year of the single rotational patient assignment as the day that was on the same day of the week and within 2 calendar days of the first day in the first year of parallel rotational patient assignment. That is, June 5th, 2017 matched with June 4th, 2018 and so on.

For the study, we considered those patients assigned to CA and CB. Patients who are sent to a specific medical specialty or arrive during the night are excluded from the study.

Interventions

During both periods, patients arrive either by their own means (normal arrivals) or in an ambulance, and in the first case, a quick administrative registration process must be carried out. Then all patients underwent nursing triage in a dedicated examination room, at which time an encounter record and patient chart were generated. The triage process classified the patients according to their severity on 5 acuity levels, assigned them to one of the two different care circuits, and within the selected circuit they assigned patients to a specific physician. This time was recorded as the triage time. Then, physicians noted the time of the first evaluation by actively “claiming” the patient on the electronic tracking board. The time at which this happened was recorded as the provider time, and this workflow as well as the triage process did not change during the study period (June 5, 2017 to July 1, 2018).

The difference between both periods is the patient assignment algorithm in CA. As previously mentioned, the assignment of patients algorithm in CB is maintained because CB is the less crowded circuit and its patients are too critical to take part in the experiment. These patients are also used as a control.

During the single rotational patient assignment period (June 5, 2017 to July 2, 2017), the triage nurses manually assigned patients to physicians within both care circuit rotationally as they arrive to the ED. During the parallel rotational patient assignment period (June 4, 2018 to July 1, 2018), within CA a software assigned patients to physicians rotationally taking into account the priority. P3 patients were considered by physicians as severe patients, and P4 and P5 patients as mild patients. Within CB, the assignment did not change from the previous period.

Measurements

Data for age; sex; ESI score; ED volume; ED volume that needs medical test; LOS; APT, APT limit exceeded; and early returns were extracted from the electronic medical record. Rates of left before being seen were not reliable because the information sheet of these patients sometimes the time of the first evaluation has been recorded without actually having been seen by the physician. This happens when physician "claims" the patient on the electronic tracking board before realizing they have already left. These indicate "false" dropouts after been seeing.

A survey was developed and conducted to investigate the ED users' (triage nurses) assessment, perception, and acceptance of the software and the new assignment rule implemented. This was approved by the nursery chief for quality and development and it was administered to all staff that had worked at triage during the pilot test period.

We defined LOS as the interval between when the patient registered in the department and when he or she is discharged. Sometimes patients are sent home, other times they are admitted to hospital and have to wait after having been discharged, etc. We defined APT as the interval between when the patient is triaged in the department and the time at which the physician claimed the patient on the electronic tracking board. We report LOS and APT in minutes. We define APT limit exceeded as not being seen by the physician within the time limit fixed by the triage system according to the patient's priority (see Table 2.2). We defined early return as returning to the ED within 72 hours of discharge. We report APT limit exceeded and early returns as a ratio. We also collected data for potential confounding variables (variables that may correlate with both the dependent and independent variables being studied). The confounding variables we identified were the patient demographics of age, sex, acuity, and need of medical, as well as the operational metrics of daily ED volume. The requirement of some medical tests for a patient involves the patient staying in the system until the results are obtained for a reevaluation by the physician, longer time in the ED. Physician staffing and nursing staffing are constant in both period studied. We measured age in integral years on the day of arrival. We assigned sex according to patient declaration. We measured acuity through the ESI score, which the nursing staff assigned in standard fashion (1 to 5). We defined daily ED volume of each circuit as the number of patients who registered in each care circuit between 8:00 and 21:00 in the day in question and total daily ED volume as the number of patients who registered in the ED between 8:00 in the day in question and 8:00 in the next morning.

Outcomes:

The primary outcome measures were APT, ratio of patients exceeding the APT limit, and LOS for patients starting their evaluation between 12:00 to 20:00. We considered the most overcrowded period of the day not to have the results influenced by physicians' behavior. Prior research suggests that servers work slower at low workloads because there is no need to work fast because of the slack capacity [159], and physicians, as strategic servers [154] can adjust they service rate by slowing down their work pace like in other sectors [175]. Other primary results are the equity of workload assigned to physicians in terms of the daily average ranges

of patients of each type assigned to physicians. Finally, secondary measures included the ratio of early return patients and the software ED users' perception as well as physicians' experience.

Analysis

Data for age; sex; ESI score; ED volume; LOS; APT, APT limit exceeded; and early returns were extracted from the electronic medical record. We report age, sex, daily ED volume as medians with interquartile range and note means and standard deviations (SDs) for comparison. We report sex as the percentage of female patients. We report ESI scores as counts and percentages for each level.

We also report total ED daily volume (from 8:00 to 21:00) in each care circuit as medians with interquartile range and note means and standard deviations (SDs) for comparison. Volumes in both periods are compered by differences in mean and median using t-test and Mann Whitney U test respectively.

Once the similarity in demand is demonstrated and knowing the staff is the same, in the primary analysis, we report ATP, and LoS as medians with interquartile ranges, and report them in minutes. We note means and SDs for comparison. We present ATP limit exceeded, and early returns as ratios. LoS and APT were compared by differences in mean and median minutes; all other metrics were compared by differences in proportions.

In the secondary analysis, we used regression models to control for patient and ED characteristics. We applied a log transformation and used a linear regression on LOS and APT to measure improvement. All other outcomes were modelled with log-binomial regression to measure relative risk. We also stratified the results for LOS by patients needing medical tests and discharged after first consultation patients. ESI score was categorized as high (3) and low (4, or 5) for all regression models.

Finally, to assess ED users' opinions of the parallel rotational system's effect on the work environment, a survey of nurses who had worked in triage under both systems was conducted one month after the change was implemented. The survey included questions about general satisfaction with the new system and perceptions of the change related with technical issues (ease of use, incident resolution, etc.).

Results

Characteristics of study subjects

In the last month of June before the implementation of the parallel rotational patient assignment, there were 9063 during 28 days (June 5 to July 2, 2017, both included): 3753 patients were assisted in CA, 2228 in care CB, and the rest of patients were assisted by specific specialist (psychiatry, ophthalmology) or during the night. In the period of this new patient assignment, there were 8892 visits during the same number of days (June 4th to July 1st, 2018):

3667 patients were assisted in CA, 2168 in CB, and the rest of patients were assisted by specific specialist (psychiatry, ophthalmology) or during the night. They all were considered for the study of workload distribution among physicians.

For the care CA time study, we considered patients who arrived to the ED during the most overcrowded period of the day, from 12:00 to 20:00: 2320 patients in 2017 and 2203 patients in 2018. We had to exclude patients' data from June 21, 2018, as the computer system did not work well and they were wrong (76 patients), majority of patients had been evaluated by the physician before the arriving to the ED.

For 38 visits in the period of singular rotational patient assignment and 33 visits in the period of parallel rotational patient assignment, data for APT were missing, and we excluded these visits from all subsequent analyses related with APT. For 13 visits in the period of singular rotational patient assignment and 14 visits in the period of rotational patient assignment, we could not reasonably determine APT and thus excluded these visits from APT (and APT limit exceeded) analysis. In all of these cases, the documented APT less than zero or greater than LOS.

For 6 visits in the period of singular rotational patient assignment and 6 visits in the period of parallel rotational patient assignment, data for LOS were missing, and we excluded these visits from all subsequent LOS analyses. For 44 visits in the period of singular rotational patient assignment and 27 visits in the period of parallel rotational patient assignment, we could not reasonably determine LOS and thus excluded these visits from LOS analysis. Reasons for exclusions included LOS less than zero, or greater than 24 hours when people leave the ED without warning.

The survey was provided to all nurses who had use the new software (those who had worked during June 2018 at triage) and it was answered by 19 of them.

Main results

We report patient characteristics (age, sex, and ESI score) in Table 4.2 and ED daily volume in Table 4.3. Physician staffing and nursing staffing are equal in both periods.

Table 4.2. Patient characteristics during both periods.

Single Rotational Patient Assignment, N=9063				**Parallel Rotational Patient Assignment, N=8892**			
ESI Score	**Patients (%)**	SEX (%) Female	Age Mean (SD)	**ESI Score**	**Patients (%)**	SEX (%) Female	Age, y Mean (SD)
*	80 (0.88%)			*	106 (1.19%)		
1	85 (0.94%)	30 (35.29%)	68.4 (17.32)	**1**	48 (0.54%)	20 (41.67%)	66.71 (15)
2	1209 (13.34%)	550 (45.49%)	62.943 (21.835)	**2**	1191 (13.39%)	534 (44.84%)	63.588 (21.183)
3	4447 (49.07%)	2258 (50.776%)	57.975 (21.283)	**3**	4414 (49.64%)	2210 (50.068%)	57.819 (21.54)
4	2969 (32.76%)	1451 (48.872%)	47.425 (18.125)	**4**	2860 (32.16%)	1418 (49.58%)	46.042 (17.952)
5	273 (3.01%)	136 (49.82%)	50.79 (19.53)	**5**	273 (3.07%)	127 (46.52%)	48.67 (18.84)

Table 4.3.ED daily volume.

ED daily volume		
P1		
Total		
Median (IQR)	3.00 (2.75)	1.50 (2.00)
Mean (SD)	3.04 (1.99)	1.71 (1.18)
Medical Test needed		
Median (IQR)	1.00 (0.19)	1.00 (0.50)
Mean (SD)	0.90 (0.18)	0.71 (0.37)
P2		
Total		
Median (IQR)	43.00 (11.50)	41.00 (5.00)
Mean (SD)	43.18 (6.92)	42.54 (7.43)
Medical Test needed		
Median (IQR)	0.81 (0.08)	0.83 (0.11)
Mean (SD)	0.81 (0.06)	0.81 (0.07)
P3		
Total		
Median (IQR)	161.50 (29.00)	158.50 (24.00)
Mean (SD)	158.82 (23.49)	157.64 (18.94)
Medical Test needed		
Median (IQR)	0.67 (0.06)	0.64 (0.06)
Mean (SD)	0.67 (0.05)	0.66 (0.04)
P4		
Total		
Median (IQR)	104.50 (16.50)	99.50 (21.50)
Mean (SD)	106.04 (11.44)	102.14 (16.65)
Medical Test needed		
Median (IQR)	0.46 (0.07)	0.45 (0.09)
Mean (SD)	0.47 (0.05)	0.46 (0.07)
P5		
Total		
Median (IQR)	9.00 (5.75)	8.50 (6.75)
Mean (SD)	9.75 (4.06)	9.75 (5.18)
Medical Test needed		
Median (IQR)	0.17 (0.14)	0.17 (0.24)
Mean (SD)	0.16 (0.11)	0.20 (0.15)
All patients		
Total		
Median (IQR)	280.50 (35.75)	270.00 (34.75)
Mean (SD)	274.61 (28.84)	269.54 (21.47)
Medical Test needed		
Median (IQR)	0.58 (0.03)	0.57 (0.05)
Mean (SD)	0.58 (0.03)	0.57 (0.04)

We also report ED daily volume according to patient's priority and possible associated medical test in each care circuit, see Table 4.4 and Table 4.5. If a patient does not need medical test, he or she is discharged after a first consultation with the physician.

Table 4.4. CA daily volume (8:00-21:00).

SIS 3	SRPA	PRPA	Difference	95% CI
Total				
Median (IQR)	62,00 (18,25)	60,50 (14,00)	-1	-7.00 to 4.00
Mean (SD)	63,29 (11,75)	61,18 (13,25)	2.11	-4.61 to 8.82
Medical Test needed				
Median (IQR)	0,6909 (0,1131)	0,7101 (0,0684)	0.00245	-0.02416 to 0.03171
Mean (SD)	0,6944 (0,0729)	0,7034 (0,0542)	-0.009	-0.0435 to 0.0255
SIS 45				
Total				
Median (IQR)	70,50 (13,25)	68,50 (17,50)	-1	-6.00 to 5.00
Mean (SD)	70,96 (8,73)	70,00 (13,29)	0.96	-5.09 to 7.01
Medical Test needed				
Median (IQR)	0,5542 (0,0833)	0,5341 (0,1069)	-0.00513	-0.03883 to 0.02779
Mean (SD)	0,5544 (0,0674)	0,5424 (0,0704)	0.012	-0.0250 to 0.0489
Aggregated				
Total				
Median (IQR)	133,50 (18,00)	128,00 (19,75)	-1	-9.00 to 5.00
Mean (SD)	134,25 (11,95)	131,18 (14,29)	3.07	-3.99 to 10.14
Medical Test needed				
Median (IQR)	0,6292 (0,1046)	0,6104 (0,0948)	-0.00177	-0.03090 to 0.02442
Mean (SD)	0,6206 (0,0611)	0,6173 (0,0567)	0.0033	-0.0283 to 0.0349

Table 4.5. CB daily volume (8:00-21:00).

SIS 1	SRPA	PRPA	Difference	95% CI
Total				
Median (IQR)	2,000 (2,750)	1,000 (1,000)	0	-1.0003 to 0.0000
Mean (SD)	1,964 (1,575)	0,964 (0,838)	1	0.319 to 1.681
Medical Test needed				
Median (IQR)	0,5000 (1,0000)	1,0000 (1,0000)	0	-0.0002 to 0.0000
Mean (SD)	0,5298 (0,3930)	0,6190 (0,4861)	-0.089	-0.326 to 0.148
SIS 2				
Total				
Median (IQR)	28,500 (7,750)	27,00 (6,75)	-1	-3.000 to 2.002
Mean (SD)	27,893 (5,252)	27,25 (6,23)	0.64	-2.45 to 3.73
Medical Test needed				
Median (IQR)	0,7289 (0,1373)	0,6809 (0,1366)	-0.0082	-0.05000 to 0.02765
Mean (SD)	0,6993 (0,0963)	0,6789 (0,0863)	0.0205	-0.0285 to 0.0695
SIS3				
Total				
Median (IQR)	51,50 (10,75)	53,00 (13,75)	0	-3.998 to 2.998
Mean (SD)	50,50 (10,38)	49,86 (8,11)	0.64	-4.36 to 5.64
Medical Test needed				
Median (IQR)	0,7196 (0,1218)	0,6988 (0,1020)	-0.01016	-0.05233 to 0.02040
Mean (SD)	0,7159 (0,0796)	0,6873 (0,0813)	0.0286	-0.0146 to 0.0717
Aggregated				
Total				
Median (IQR)	80,00 (15,50)	79,00 (17,00)	-1	-6.001 to 3.997
Mean (SD)	80,36 (12,25)	78,07 (9,88)	2.29	-3.68 to 8.26
Medical Test needed				
Median (IQR)	0,7203 (0,1018)	0,6759 (0,0700)	-0.01199	-0.04152 to 0.01423
Mean (SD)	0,7086 (0,0672)	0,6860 (0,0641)	0.0226	-0.0126 to 0.0578

We report unadjusted outcomes related to patient quality of care in Table 4.6. During the period of parallel rotational patient assignment, LOS, APT, and rate of patients exceeding their APT limit were all lower. There were no significant changes with respect to early returns.

Table 4.6. Unadjusted patient outcomes.

APT. min					
P3	SRPA	PRPA	Difference	95% IC	T-test p-value
Median (IQR)	58.00 (54.25)	47.000 (49.00)	8	5.002 to 11.998	<0.001
Mean (SD)	65.77 (45.59)	56.80 (43.08)	8.96	5.16 to 12.76	<0.001
P4	SRPA	PRPA	Difference	95% IC	T-test p-value
Median (IQR)	70.00 (83.50)	64.00 (75.00)	5	0.999 to 8.997	0.0124
Mean (SD)	80.68 (55.78)	74.53 (52.58)	6.15	1.69 to 10.62	0.007
APT exceeded. min					
P3	SRPA	PRPA	Difference	95% IC	T-test p-value
Median (IQR)	0.00 (1.00)	0.00 (1.00)	0	-0.0000 to 0.0000	<0.001
Mean (SD)	0.4849 (0.5000)	0.3752 (0.4844)	0.117996	0.0759340 to 0.160058	<0.001
P4	SRPA	PRPA	Difference	95% IC	T-test p-value
Median (IQR)	0.00 (0.00)	0.00 (0.00)	0	-0.0000 to 0.0000	0.0196
Mean (SD)	0.2257 (0.4182)	0.1860 (0.3893)	0.0397106	0.00646501 to 0.0729562	0.019
LOS, min					
P3	SRPA	PRPA	Difference	95% IC	Mann-Whitney //T-test p-value
Median (IQR)	191.00 (165.25)	169.00 (149.00)	23	14 to 33	<0.001
Mean (SD)	209.28 (114.12)	184.27 (107.12)	25.02	15.49 to 34.55	<0.001
P45					
Median (IQR)	158.50 (144.75)	143.00 (137.00)	13	5 to 21	0.0014
Mean (SD)	176.72 (109.35)	162.18 (102.96)	14.54	5.85 to 23.23	0.001
Early Return (72h)					
Total	SRPA	PRPA	Difference	95% IC	T-test/Z-test p-value
P3					
Median (IQR)	0.00 (0.00)	0.00 (0.00)	-0.0	0 to -0	0.446
Mean (SD)	0.0183 (0.134)	0.014 (0.118)	0.004	-0.006 to 0.015	0.442
P45					
Median (IQR)	0.00 (0.00)	0.00 (0.00)	-0.0	0 to -0	0.184
Mean (SD)	0.01704 (0.12946)	0.01056 (0.10228)	0.00647	-0.00302 to 0.01597	0.181
Total					
Median (IQR)	0.00 (0.00)	0.00 (0.00)	-0.0	0 to -0	0.136
Mean (SD)	0.018 (0.132)	0.012 (0.110)	0.005	-0.002 to 0.013	0.133

Figure 4.9 represents the interval plot for the APT and LOS for priority 3 and priority 4&5 during the period of the single rotational patient assignment and during the parallel rotational patient assignment. After the implementation of the new assignment rule, the time limits for first consultation for both type of patients (60 minutes and 120 minutes respectively) are achieved.

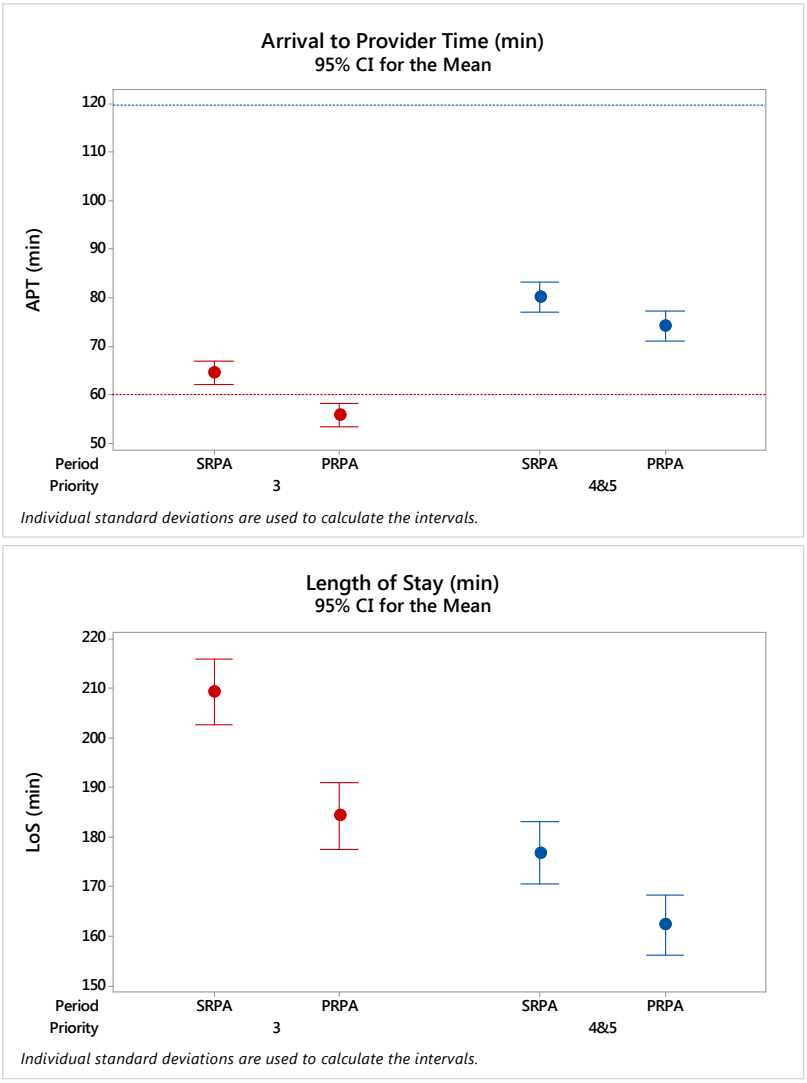


Figure 4.9. LOS and APT for priority 3 and priority 4&5 patients during both periods. The horizontal broken line represent the APT limit for each priority.

We report the variability of the workload assigned to each physicians in Table 4.7. It is represented the daily average range of patients of each priority assigned to physicians. All patients' types' ranges are significantly lower during the period of parallel rotational patient assignment as this new assignment rule tries to balance them among physicians.

Table 4.7. Physician Outcomes: range of the number of patients of each type assigned to different physicians.

Range among physicians					
Priority 3	SRPA	PRPA	Difference	95% IC	T-test p-value
Median (IQR)	3.85 (1.73)	1.97 (0.89)	1.695	1.158 to 2.242	<0.001
Mean (SD)	3.80 (1.00)	2.12 (0.71)	1,680	1,214 to 2,146	<0.001
Priority 4&5					
Median (IQR)	4.17 (1.30)	2.04 (1.67)	2.015	1.434 to 2.623	<0.001
Mean (SD)	4.42 (1.56)	2.25 (0.97)	2,179	1,481 to 2,877	<0.001

We report our regression analysis results in Table 4.8. Regression analysis confirmed that parallel rotational patient assignment was associated with a decrease in APT, APT limit exceeded, and LOS for all type of patients. As it is expected, in case of the LOS, apart from patient priority assigned in triage, ED volume, and patient assignment rule, the patient characteristics (need of medical tests, age, etc.) are also significant.

Table 4.8. Regression analysis for patients outcomes

	KPIs					
	APT		APT limit exceeded		LOS	
Term	Coef	p-Value	Coef	p-Value	Coef	p-Value
Constant	2,358	<0,001	-3,882	<0,001	4,169	<0,001
Priority						
4&5	0,2147	<0,001	-1,0751	<0,001	-0,139	<0,001
Patient Assignment						
PRPA	-0,097	<0,001	-0,340	<0,001	-0,098	<0,001
CA Daily Patient Arrival	0,011	<0,001	0,0278	<0,001	0,005	<0,001
Patient Characteristics						
Age	0,001	0,429	0,001	0,590	0,002	<0,001
Sex						
Male	0,002	0,949	-0,038	0,578	-0,040	0,045
Medical test needed						
YES	-0,009	0,727	-0,042	0,556	4,169	<0,001

Regression analysis confirmed that parallel rotational patient assignment was associated with a decrease in daily average range of patients of each priority assigned to physicians. Other factors such as ED volume are not significant, see Table 4.9.

Table 4.9. Regression analysis for the number of patients of each type distribution range (physicians' outcomes).

	High priority Patients (P3)		Low priority Patients (P4&5)	
Term	Coef	p-Value	Coef	p-Value
Constant	4,60	0,000	4,23	0,013
CA Daily Patient Arrival	-0,00668	0,407	0,0095	0,423
P3 CA Daily Patient Arrival	-0,0043	0,755	-0,0151	0,461
Patient Assignment				
PRPA	-1,691	<0,001	-2,241	<0,001

All respondents of the survey agree that they recommend to definitively use the new assignment rule in CA where the pilot test took place and extent it to CB. Moreover, the ED physicians experienced positive effects of new parallel rotational patient assignment and considered it fairer than the previous one.

Limitations

We report findings of correlation and not causation given that we present before-and-after data. This was an observational study, and, although our regression analysis attempts to account for multiple confounding variables, such an attempt does not guarantee that all key factors were

incorporated into the model. Our core physician group underwent little change and no other systematic changes in the ED delivery of care were undertaken at this facility between these 2 periods, however other unidentified factors such as physician turnover, the Hawthorne effect, etc. may have contributed to the improvement in patient waiting times, and LOS.

Thus, the use of a comparison group, more critical patients care circuit (CB), was also designed to control for the extensive external factors such as laboratory turn-around time, other hospital support services, the admission process, and changes in the managed care environment, all of which may affect LOS and may have varied during the year of the study. The effect of these variables should have been similar for both groups. However, as it will be mentioned in the Discussion Section, there were no changes and the performance indicators for the control group (CB) remained constant.

We measure patient quality of care considering waiting time and LOS, and physicians' satisfaction or quality considering the reduction in the workload assigned variability among them and report anecdotal physician sentiment in our discussion. It is possible that a methodical assessment of physician attitudes – similar to the surveys developed to triage nurses – would reveal objective results that were not apparent in subjective interviews.

We do not believe that staff altered their practice with a goal of showing improvement with parallel rotational patient assignment, however, we cannot rigorously exclude this possibility. Also, even if we relied on systems-generated data, we audited all collected recorded data and removed illogical values, which raises the possibility that other data, although logical, were imperfect. This is an inherent problem with any study that relies on large amounts of data and as there was no change in electronic data processing during the study period, we also believe that if imperfections happens, it is unlikely that there was an unequal distribution of them between the 2 groups.

4.3.4 Conclusion

Our study found that a patient assignment system taking into account patients priority (high or low in this study), what we call a parallel rotation, was associated with reduced patient waiting times, reduced ratio of patients exceeding the APT limit, reduced LOS from patients point of view, and reduced difference in the number of patients of each type assigned to each physicians from the medical staff points of view. This assignment system may serve as a useful model that many EDs can implement to improve patient care and ED throughput as well as medical staff satisfaction.

A previous work related to this topic compared the rotational patient assignment with other front-end processes designed to improve patient flow, the physician in triage [65] obtaining no statistically significant differences. Other investigations of different groups studied the transition from physician self-assignment to rotational patients assignments –single rotational

assignments – , which was associated with significant improvements in some ED operational metrics (some of them were carried out before the new-universal adoption of electronic patient tracking [135], [136], [149], [151], and others more recently [140]). In three of the them patients were assigned to alternating teams reporting an increase in patient satisfaction [135] [136], a decrease in APT [135], [149], while in others patients were assigned to physicians rotationally reporting a decrease in APT and LOS [140]. An older paper described a semicontrolled study in which rotational patient assignment was instituted for residents on the "medical side" of a "medical side/surgical side" ED at a teaching facility. It reported that LOS improved on the medical side by 15%, whereas it increased on the surgical side [151]. Similar to this study, in our facility the change was introduced in the less critical patients care circuit (CA) while the most critical patients care circuit (CB) maintained the previous patient assignment rule. However, in our study several ED operational metrics in the change implemented circuit improved while the other circuit was not negatively affected and remained operating in the same way as before with no significant performance changes.

All previous studies demonstrated that opposite to queuing theory, the change from self-assigning patients to the inflexible system of rotational assigning patient to physicians leads to increases in efficiency although it seems counterintuitive. The gain in responsibility [135], [136], patients ownership [150], [155], [162], equitable distribution of the number of patients [135], [162], [176] had been reported as reasons for that results. And this system has been reported until now as the best to be used.

This singular rotational assignment system ensure an equal distribution of the number of patients, however, day-to-day (or even patient-to-patient) randomness can lead to perceptions of workload inequality because the average acuity of patients assigned to one physician might be higher than that of patients assigned to another. The medical staff of the HCN as well as physicians in other studies, who has reported this system as "mercilessly fair" [140], has complain about this issue [151]. It also affects patients' quality of care by assigning more high priority patients to the same physicians slowing the attention of their patients while other physicians are almost idle because they were assigned all low severity patients.

Our intervention go further and prove superior to previous ones as it uses a more advanced assignment algorithm taking into account patient ESI score. This algorithms are commonly used to solve similar problems in other industries such as routing incoming calls in call centers [158], [177].

This solution apart from improving the LOS, ratio of patients exceeding the APT limit, and APT for all types of patients from patients point of view, it also improves the physicians' satisfaction as it was expressed by them in terms of equity in the number of patients of each type assigned to each physician. Objectively, the range of high priority patients and low priority patients among them was reduced and subjectively, they had a positive sense of workload

similarity among physicians. They stated that they could cope better with their patients as it was easier to manage patients from different urgencies at the same time.

Our intervention does not rely on incremental resources for success; parallel rotational patient assignment can be instituted without additional nurses, physicians, or space. Moreover, subjective nursing response to this intervention has been positive, they all answered in the survey that they prefer this fairer system and they found it very user-friendly and easy.

Finally, a parallel rotational patient assignment system is almost assuredly appropriate for those in which allocating patients to physicians also works. We believe that this intervention was successful at our institution as physicians in public EDs are salaried, and the financial incentive is lacking. In a fee-for-service model, seeing more patients may be incentivized by increasing billing reimbursements. Furthermore, the simulation results as well as the model validity were demonstrated in reality.

4.4 Preliminary assessment and work in progress of stress based policies

Preliminary results

In this section, preliminary results of the stress based policy are shown. This rule cannot be implemented in the ED of the HCN due to the information system. It is not possible to access information of pending patients in real time so job stress score cannot be calculated.

We compare the stress based policy with the one currently used in some EDs including the HCN, and considered to be the best in literature, the SRR. These results demonstrate superiority of stress based policy especially in the equal distribution of the instant workload and the reduction of the waiting time, APT.

Figure 3.11 of Section 3.6 of Chapter 3 represents the instantaneous real job stress level experienced by the different physicians during their work shift by using the current assignment rule (SRR). It shows the workload inequities among physicians, which were expected as patients are assigned to a specific physician rotationally as they are triaged without considering their complexity, their priority, or the physician’s pending patients.

Figure 4.10 represents the job stress of the six physicians the same Monday of Figure 3.11 by using the stress based policy assignment rule during the work shift. It reproduces the same demand and patients’ characteristics, and reorganizing events considering that patients were assigned to the physicians according to this new proposed criterion.

Figure 4.10. Stress associated to each physician along a specific work shift by using a DSS based on the minimum stress score assignment rule.

A sample of 50 days has been considered (including that represented in Figure 3.11 and Figure 4.10) and as a result, the variability in the instantaneous job stress experienced by physicians during the workshift has decreased in more than 60%. However, variability reduction is not the only benefit related to job stress, by balancing job stress during the workshift, the average of stress per physician has been reduced in 10% when using the new proposed assignment rule based on the stress score (see Figure 4.11). The job stress score experienced by a physicians developed in Chapter 3 is not a linear function of the number of patients assigned to them but it is a function of uncertainty associated to each of them as well as time pressure. In this case, a better distribution of patients may reduce the patients exceeding the limit and avoid the accumulation of patients in this situation in the same physician's queue. This reduces the total job stress score of the ED.

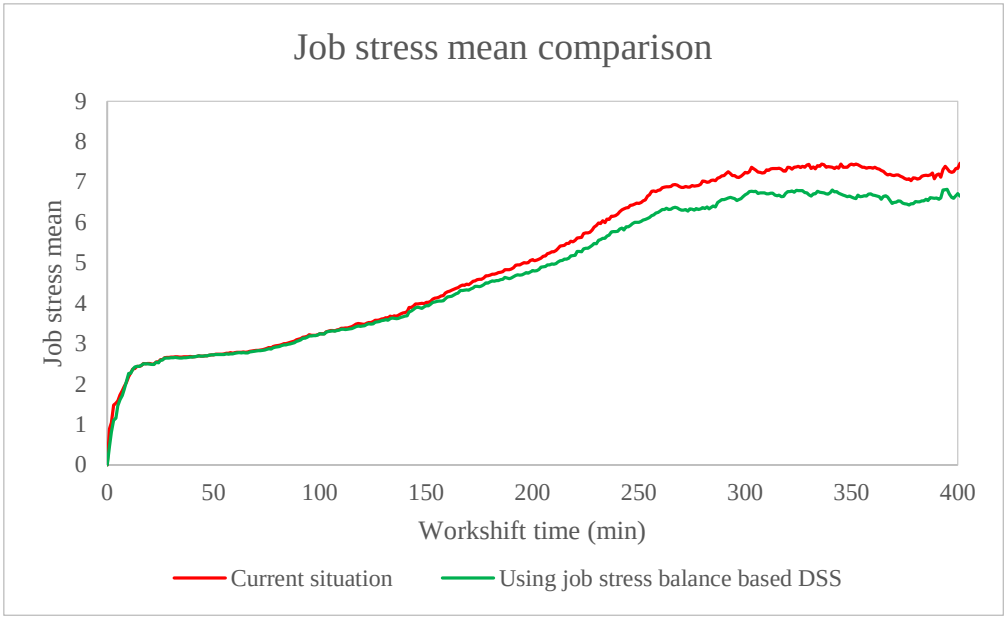


Figure 4.11. Average stress per physicians during a work shift by using the current rule and the new job stress balance rule.

Moreover, KPIs such as ATP of patients has been taken into account, which is significantly reduced for all patients. It has been improved in 3% for high severity patients (P3) and in 10% for low severity patients (P45). These improvements also have a significant impact on the percentage of patients who exceed the ATP time limit, which is reduced in 12.2% and 23.2% for high severity patients and low severity patient respectively compared to the SRR results (see Figure 4.12)

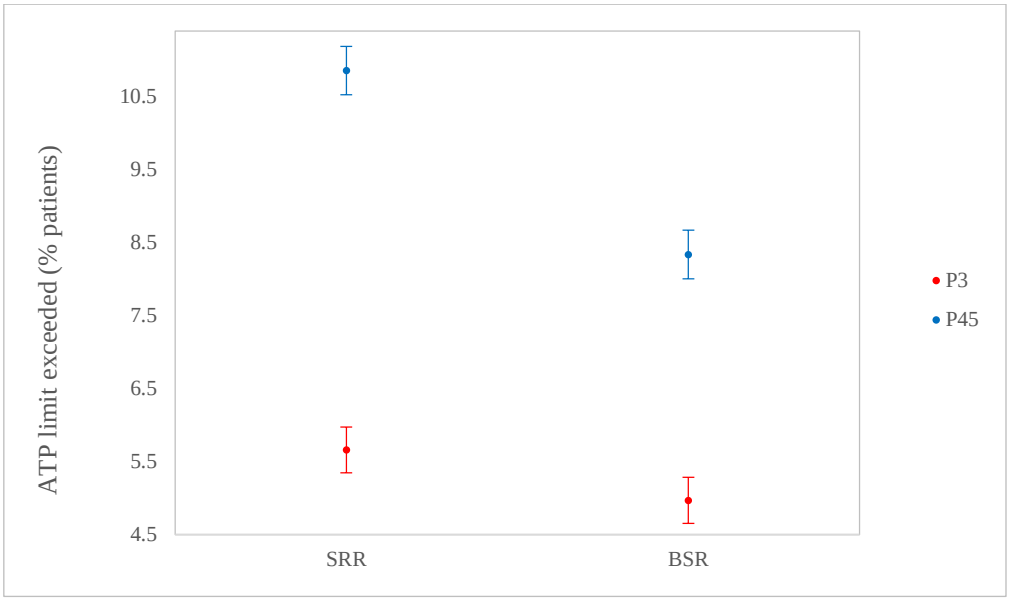


Figure 4.12. Percentage of patients who exceed their ATP time limit.

Work in progress

As described in Section 4.2.3, the stress balancing rule, SBR, has been criticized by physicians who stated that they do not work at the same rate, which may cause inequities in the total workload assigned to each physician throughout the entire workshift. This rule does not take into consideration the historical data, that is, the patients they have assisted from the beginning of the workshift until the moment of being allocated new patient. It only considers the instantaneous situation when assigning a patient to a physician, the job stress they are experiencing due to pending workload.

In this sense, rotational rules, such as the SRR and the MRR, equitably distribute workload throughout the entire period (workshift) across physicians while the SBR tries to equitably distribute the job stress due to workload across physicians at every instance.

To solve this problem, the physicians' job stress and completed workload balancing rule (SWBR) was proposed, which considers not only pending patients at that moment but also already discharged patients. That is, the labor developed by each physician from the beginning of the workshift until the moment a new patient arrives. As with the SBR, the SWBR's implementation is not currently possible in the ED for lack of an appropriate information system that provides real time information to be used to calculate the patients' assignment criterion.

Currently we are investigating the influence of λ parameter for linearly combine job stress, $Y_k(t)$, and completed workload, $C_k(t)$. The higher λ is, $\lambda \in [0,1]$, the shorter APT, average job stress, and job stress variability across physicians are. SBR is a particularization of the SWBR when $\lambda = 1$, which gets the shortest value for the just mentioned KPIs (see Figure 4.13).

Currently we are investigating the influence of λ parameter for linearly combine job stress, $Y_k(t)$, and completed workload, $C_k(t)$ in order to decide the "optimal" value. The higher λ is, $\lambda \in [0,1]$, the shorter APT, average job stress, and job stress variability across physicians are. SBR is a particularization of the SWBR when $\lambda = 1$, which gets the shortest value for the just mentioned KPIs (see Figure 4.13). Meanwhile, it provides the worst results for variability of the completed workload across physicians at the end of the workshift.

We want to attain more than one goal in selecting the assignment rule criterion: optimizing not only patient's quality of care but also physicians work conditions so the decision of the λ involves multiple and conflicting objectives and should be treated in a multiobjective framework.

Traditionally, one approach consists in using weights to combine the objective functions together. However, in most cases there is not enough information to establish the relative importance among objectives that leads to an "optimal" solution. Our approach is the Pareto curve, which applies to bicriteria optimization problems and can be used to multi-criteria

problems considering two objectives at the same time. It graphically illustrates the trade-off decision between objectives facilitating the decision process.

As an example, Figure 4.13 represents the optimal trade-off in the space spanned by two axes representing APT (patients care quality and satisfaction contribution) and variability of workload developed by the end of the work shift across physicians (working conditions contribution) respectively. It gives a rigorous yardstick for measuring the performance of the ED system by using the possible available alternatives ($\lambda \in [0,1]$) by tracing the optimal trade-off between these two competing aims. This is a useful tool to formulate a preference facilitating the election of the λ "optimal value" by deciding where on the Pareto curve the "optimal" solution lies. This approximated curve has been calculated by sampling various points in the Pareto curve but needs to be investigated and more exact calculated.

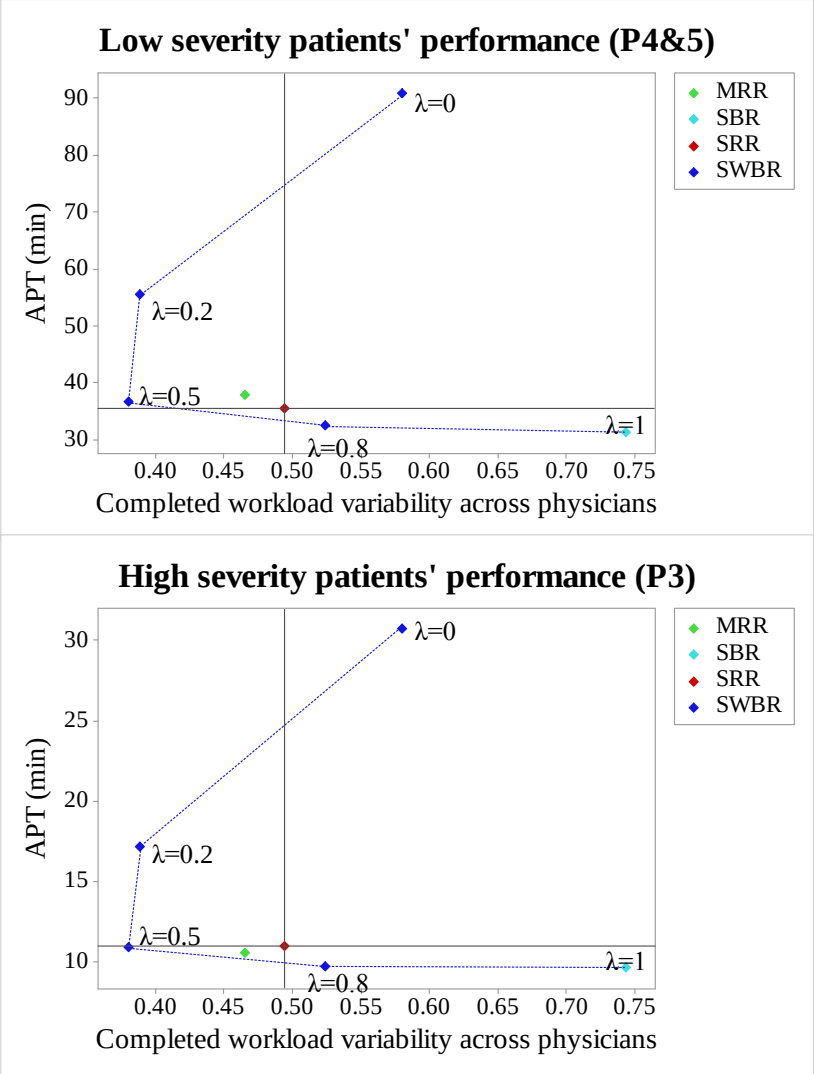


Figure 4.13 APT and Completed Workload Variability by using different PPAR. There is a sample of 5 values of λ for the parametric rule SWBR.

Chapter 5 Patient flow management during treatment

5.1 Introduction

There are two phases of sequencing decisions, as explained in the introduction of this part of the thesis,: 1) determination of the order/priority in which patients are initially taken from the waiting area to start treatment with a physician, which has been addressed in previous Chapter 4, and 2) determination of the order in which patients are seen once they are under the responsibility of a physician This chapter studies the sequencing decisions of this latter phase, which are usually made by individual physicians by choosing among patients assigned to them, who are of different priorities and in different treatment stages. It has been observed wide variance in the sequencing logic of individual physicians working within the same ED [71].

Upon arrival, as mentioned in the introduction of Part I, patients undergo an initial assessment, i.e., triage, whose aim is to stratify them by illness acuity and prioritize them accordingly ([59]). Triage systems may include performance goals in terms of the percentage of patients who should have access to the physician consultation before certain time limits and should have a different time limit and a different percentage for each type of triage-level (see Table 2.2). However, most triage systems do not provide explicit guidelines on how to manage the patient flow within and among the, usually five, assigned triage levels. ED managers and physicians, motivated by the achievement of such goals, follow pre-determined rules, such as FCFS, within the same triage level and strictly follow priority across different triage levels. Nevertheless, very often, especially in cases of overcrowding, they have to use their own discretion in making patient-routing decisions, as mentioned in [60]. In this paper, by using patient-level ED visit data, the authors carried out an empirical study to understand how decision makers manage patients in the ED. They concluded that, generally, higher triage-level patients receive priority over low triage-level patients, but a lower triage-level patient who has waited longer can be prioritized over a higher triage-level patient who has waited less time. Then, they highlighted the need to consider not just the triage level but also the actual wait time in routing decisions. Therefore, the behaviour of these patient flow managers fits the so-called accumulative priority queue (APQ) policies, a term introduced by Stanford et al. [178]. Following this APQ strategy,

patients accumulate “priority points” as they wait for treatment, and the patient with the most priority points is selected when a physician finishes a service. The accumulation rate of priority points depends on the patient’s triage level. In addition, in [60], it is also concluded that patients who have waited past the target set by the triage system (for example, 30 minutes for patients level 3 in the CTAS triage system; see Table 2.2) may not receive extra priority. The ED decision makers’ behaviour is described by a two piecewise linear concave marginal waiting cost function for each triage level, in which the break point is located around the target wait times. This important empirical observation suggests a modification of the classical APQ policy to define the new APQ-h (accumulative priority queue with a finite horizon) policy, which linearly accumulates priority points while the patient is waiting until the target wait time is reached, and then, no more priority points are accumulated.

The implementation of priority strategies in a real ED needs to consider not only the prioritization of patients to access to their first physician consultation but also the management of patients already in the process of being treated. After the first consultation with a physician, some patients are discharged from the ED, while others require some clinical tests and, once the results are obtained, have a second consultation with a physician. Thus, patient management should consider the following two components of the patient flow: first, the patients arriving from triage that must be served within time-deadlines and, second, the patients already being treated, both of which have a significant feedback constituent that produces operational congestion. Therefore, when a physician becomes idle, a decision has to be made regarding whether the next patient to be seen is having their first or for a second consultation; that is, managing the portfolio of pending patients must consider both the severity of the condition and the stage of their treatment.

The consideration of conflicting objectives is typical in the analysis of healthcare systems, as in all public services in which cost objectives compete with service quality objectives. Even in a case with fixed resources, as in the case of determining operative rules for optimally managing the ED patient flow, there are several conflicting objectives that guide the measurement of management performance. One of the main ED performance measures is the arrival to provider time (APT) (“door to doc”), which is defined as the interval between the time a patient arrives at the ED and the time an attending physician sees the patient [73]. Another important objective is minimizing the length of stay (LoS) in the ED. As was stated before, the upper limit for the APT is set for each type of patient (see Table 2.2) but can also be imposed onto the other performance measures; for example, EDs in hospitals in the United Kingdom should complete and discharge 98% of patients within 4 hours, as it is mandated by the government (Mayhew and Smith [179]). The patient-flow management strategy should be selected to accomplish the goals and optimize the objectives set by the hospital direction board. One main characteristic of the APQ and APQ-h management policies is their capacity to represent very different dynamic priority rules by changing the value of the rates at which the different types of patients accumulate priority.

The main aim of this chapter is to explore the implementation of APQ-h managing policy in a real ED framework that considers the acuity level of patients, the stage of treatment, the stochasticity of the ED and the different objectives set by managers. Specifically, the main contributions of this chapter are as follows:

- The analysis of the ED patient flow management problem in a setting not previously considered in the literature to include the different acuity levels of the patients, several stages for treatment, the stochastic environment in which EDs evolve and different Key Performance Indicators (KPIs).
- The proposal of the APQ-h policy to represent the real patient-routing decision making observed in empirical studies.
- The definition of a multi-objective and stochastic optimization problem to obtain the optimal APQ-h policy which is solved by a simulation-based optimization method.
- Testing the performance of APQ-h policies by using a simulation model that reproduces the main features of a real ED, i.e., the stochasticity in arrivals, service times and paths thorough the ED.
- A sensitivity analysis to determine the influence of the optimal APQ-h policy on the structure and on the ED's KPI of factors, such as the variability in the patient arrival pattern, the mix of patients, and the congestion level.
- Comparing the performance of the APQ-h with pure priority disciplines, to show its superiority. It also outperforms the priority rule used in the ED of the Hospital Compound of Navarre.

The rest of the chapter is organized as follows. In Section 5.2, the related literature is reviewed. The characteristics of the ED patient flow and its KPIs are presented in Subsections 5.3.1 and 5.3.2, respectively. Section 5.3.3 presents the management policies considered in this chapter, the pure priority disciplines and the APQ-h discipline, as well as simulation-based optimization methodology to determine the optimal management policy. Section 5.4 is focused on the case study, in which the main features of a real ED and the simulation model are described, and then, the optimal APQ-h policy is obtained and its performance is compared with the pure priority disciplines, including the currently followed by the majority of physicians in the ED studied. The last Subsection includes a sensitivity analysis on the weights of the objective function. Finally, Section 5.5 presents the results and conclusions from an extended computational analysis carried out to test the pure priority disciplines and optimal APQ-h disciplines in a variety of EDs defined by different occupancy ratios, patient arrival patterns and mixes of patients. We end the chapter with a conclusion Section. Finally, Appendix F includes a table with all acronyms used in this chapter and Appendix H additional numerical results.

5.2 Related literature

In most healthcare settings with no appointment system, the queue discipline is either a first-in–first-out (FIFO) or a priority discipline, depending on the acuity of the patient's illness. This priority discipline applies in the ED, where, in general, patients with life-threatening injuries are treated before others. The use of a priority discipline with a FIFO rule inside each class of patients is almost generalized in the analysis of EDs by queuing models and/or discrete event simulations (see, for example, Taylor et al. [180], Haussman [181], Siddharthan and Jones [182], Laskowski et al. [183], Mokaddis et al. [184]). It is worth mentioning the paper by McQuarrie [185] that applies the shortest processing time rule, which is known to minimize waiting times. Although routinely applied in the manufacturing context, it is difficult to justify the use of this dispatching rule in EDs, given its unfairness to the more injured patients and the added difficulty of estimating the treatment times accurately. Nevertheless, this research raises the question of using other queue disciplines than the pure priority discipline to manage ED patient flow. The dispatching rules applied in manufacturing prioritize all jobs waiting for processing on a machine (the classic paper of Panwalkar and Iskander [186] presented a summary of 113 dispatching rules). The same idea can be applied in the ED patient flow management problem, i.e., whenever a physician has finished a patient's service, the dispatching rule selects the patient with the highest priority. Dispatching rules and other prioritizing policies to manage the patient flow in an ED are usually analysed by using queuing theory models or simulation models (or both in combination).

The paper by Armory et al. [187] provides a deep queueing-network view of patient flow in hospitals, with a special focus on EDs and the in wards patient flow, as the natural way for studying and improving its performance. They pointed out how the patient flow within the ED has been widely investigated, both academically (Hall et al., [188]; Saghafian, Austin and Traub, [15]; Zeltyn et al., [189]) and in practice (IHI, [190]; McHugh et al., [58]). Among all these studies, we highlight the paper of Huang et al. [191], which addresses many of the complexities of EDs that are often ignored in queueing models; this study considers the patient triage level and the feedback of patients after the first consultation. They obtained an asymptotically optimal patient flow policy that is based on the c□ dispatching rule, which minimizes congestion costs subject to deadline constraints for the first consultation. Their analysis extended the results of Smith [192] and set the optimality of the known as the c□rule, which prioritizes among the queues of the different categories of patients and then uses the FCFS discipline inside each queue. The waiting cost was assumed to be a linear function of the sojourn time. Later, the paper of Van Mieghem [193] shows that the generalized c□ rule minimizes the average waiting costs under the heavy-traffic asymptotic regime and the cumulative holding cost is a non-decreasing convex function. Mandelbaum and Stolyar [194] and Gurvich and Whitt [195] studied the queue-length version of the Generalized c□-rule, in which the holding cost is a function of the queue length instead of the sojourn time. The aforementioned paper of Huang et al. [191] is the first to consider feedback and deadlines

simultaneously; however, the need to assume a stationary heavy traffic and the use of diffusion approximations to obtain the results do not guarantee the optimality of the proposed control rules in a real setting (for example, it is necessary to assume that during the sojourn time of a patient within the ED, the various queue lengths do not change significantly, and the service duration is negligible relative to the queueing time).

Other types of queuing models developed to study the ED patient flow optimization problem without the need of asymptotic assumptions are those that specifically analyse the APQ strategies (Stanford et al. [178]). The APQ model can be seen as a dynamic priority discipline in which patients of lower priority classes can overcome the priority of higher classes as their waiting time increases. In this way, they seek to overcome the drawback of pure disciplines that in periods of high demand, patients of the lowest priority can be "forgotten" in the system for long time periods. Kleinrock [196] obtained results about the mean waiting time before receiving service, which were extended by obtaining the waiting time distribution for each priority class in the single server and in the multi-server settings [178], [197]. All these models assume Markovian distributions (Poisson arrivals and exponentially distributed service times), and in addition, there is only one stage for the service without feedback.

The ED healthcare process can be represented by a queueing system with feedback to model the patients who need clinical tests after the first assessment and need to return for a second consultation. [60] models the ED in which patients are waiting to see a physician as a multi-class queueing system and investigates how decision makers choose which patient is the next to be seen by an available physician. They obtained strong evidence of the practical use of a sophisticated prioritization behaviour that is consistent with the APQ-h discipline and that, consequently, supports the research carried out in this chapter. Nevertheless, they only consider the prioritization to the first consultation without addressing the feedback of patients already in the process of being treated.

A related research is exposed in the paper of Ferrand et al. [198], where the patient flow management problem is analysed by using a simulation model that reproduces a real life setting that includes different acuity levels, and the stochastic environment. They conclude that dynamic priority queues outperform other approaches based on different implementations of fast tracks for low priority patients. The main difference between the Ferrand et al.'s model [198] and ours is that we consider deadline constraints for the first consultation, whose fulfilment becomes an important goal in addition to the minimization of the LoS (the only one considered in Ferrand et al. [198]). As a consequence, we do not assume the policy of prioritizing treatment over the first consultation as they do, and the management problem is addressed from a bi-objective point of view. In Zayas-Caban et al. [199] the prioritization of treatment is also criticized in a patient management problem focused on maximizing the profit when a reward is obtained from patients that complete the treatment and there could be abandonment of patients before the treatment is complete.

5.3 Physician's queue of pending patients management

5.3.1 Patient routing

The flowchart of a patient being processed through an ED is explained in detail in the introduction of this part of the thesis (see Figure 2.8). Patients arriving by ambulance or and walk-in patients (after registration) undergo a triage process, which classify them by category (level of urgency). As mentioned in the Introduction section of this part of the thesis, we consider that the triage classifies ED patients on 5 acuity levels, as is the case of CTAS (Table 2.2: Access time is the upper limit for the arrival to provider time, and performance level is the minimum percentage of patients that should satisfy the access time requirement).

Once patients are triage, they start treatment in a care circuit. Usually, EDs organize the patient care into two different care circuits, one for the more critical patients and another for less critical patients. In this chapter, we focus on the patient flow management in the less critical patient circuit, which represents the biggest volume of the patients and is the most overcrowded. However, as each circuit has dedicated resources such as physicians, nurses, waiting rooms, exploration rooms, etc. and both operate as two independent EDs under the same management policies, the following approach can be extended to other care circuits.

During treatment, each patients is evaluated by a physician, who may order clinical tests, such as blood tests, X-rays, scans, or a specialist consultation and reevaluate the patient and test results before allowing discharge. Thus, all patients initially wait in a queue for the first consultation (red arrow in Figure 2.8) in which an initial assessment can result in discharging the patient from the ED or in ordering some clinical tests, such as blood tests, X-ray, scan, specialist's consultation, etc. Once the tests and complementary diagnosis are carried out and their results are ready, the patient re-enters the queue (blue broken line arrow in Figure 2.8) and waits for a second consultation with the ED physician to be reviewed before being discharged from the ED.

After concluding a consultation, a physician has to choose a pending patient from the queue to provide a medical consultation. This queue is formed by patients of different priorities, and within these priority categories, patients can be classified into one of the following two categories: new patients who have arrived just after triage or patients who have re-entered the queue for being re-evaluated. The queue discipline implemented by the physician greatly influences the quality of the service measures, which are discussed in the next sections.

5.3.2 Key performance indicators

Assuring quality care in the ED requires the development of indicators that are valid, relevant, and feasible [200]. Welch et al. [72] and Welch et al. [73] list various metrics by which ED performance can be measured, such as the arrival to provider time (APT, or "door-to-doc"

time). This important time interval is widely used in emergency healthcare services, since many illnesses are time-dependent, and a delay in the diagnostic evaluation by a qualified medical provider could be a health risk for the patient. Most EDs define a maximum waiting time for each acuity level and set performance goals related to them, as is explained in Table 2.2's CTAS; for example, class 1 patients, the most urgent, should be immediately seen by the physician, while nonurgent patients can wait up to 120 minutes. The ratio of patients whose APT exceeds the time limit is considered a KPI.

There are also other important measures influenced by the patient flow management policies, such as the arrival to discharge time, called the "length of stay" (LoS), which has an impact on the patient's quality perception of the received healthcare service. The LoS depends directly on the treatment needed. It will be, in general, much greater for patients who need additional diagnostic tests and a second consultation than for patients who are discharged after the first consultation. The waiting times for the first and the second consultation of a patient with acuity level i are denoted by τ_i and ν_i, respectively. Thus, the total waiting time (TWT) for a patient of acuity level i is τ_i when only one consultation is needed and $\tau_i + \nu_i$ when two consultations are needed. Because the queue discipline implemented to manage the physician waiting room directly impacts those waiting times, the TWT is considered a KPI in this study.

Other KPIs could be considered, such as measuring the "overcrowding" level, which affects the availability of resources and causes an increase in the infection probability, the physician's stress level, waiting times, LoS [201], medical error probability [202], and the patient's perception of quality. Overcrowding occurs when demand exceeds available capacity, i.e., when there is no space left to meet the timely needs of the next patient requiring emergency care; however, according to [203], "No measure is universally applicable as a marker of overcrowding and should be used with caution when comparing performance between institutions". One scoring system that has become a national standard in the United States is the National ED Overcrowding Scale ("NEDOCS", http://www.nedocs.org), whose elements include total patients in the ED, as well as the waiting time of the longest admitted patient, among others. Other studies, such as that of Weiss et al. [204], which found, using multivariate regression analysis, that the combination of patients in the waiting room and the total registered patients was a better model than the NEDOCS score for quantifying paediatric ED overcrowding. Little's formula relates the average number of patients in the waiting room with the average waiting time. Thus, aiming at the minimization of the TWT implies reducing the number of patients in the ED's waiting room, which is a main contributor to overcrowding.

5.3.3 Patient flow management policies

A patient flow management policy is a rule that determines which patient will be the attended next by a physician when he/she becomes available (after ending a consultation). The implemented policies should be designed to achieve good ED performance, which is assessed by a set of KPIs, such as those defined in the previous section.

As mentioned, the physician's queue of pending patients is formed by several categories of patients, which are defined by both the illness acuity level and the healthcare service stage. The patients waiting in the physician consultation waiting room can be in one of two stages, i.e., waiting for the first consultation stage, denoted by **1C**, or waiting for re-evaluation after having some medical test, denoted by **2C**. Without loss of generality, we assume that the patients are classified into three different levels of priority according to their illness acuity as follows: high priority, denoted by **HP**; medium priority, denoted by **MP**; and low priority, denoted by **LP**. Subsequent analysis can be readily adapted to any number of acuity level categories. Therefore, the patients in the physician waiting room can be classified in one of the six categories represented in Figure 5.1, which are denoted by **1C-HP**, **1C-MP**, **1C-LP** (high-, medium-, and low-priority patients waiting for the first consultation) and **2C-HP**, **2C-MP**, **2C-LP** (high-, medium-, and low-priority patients waiting for the second consultation).

In the next subsections, policies based on pure priority disciplines and on accumulating priority queues are described.

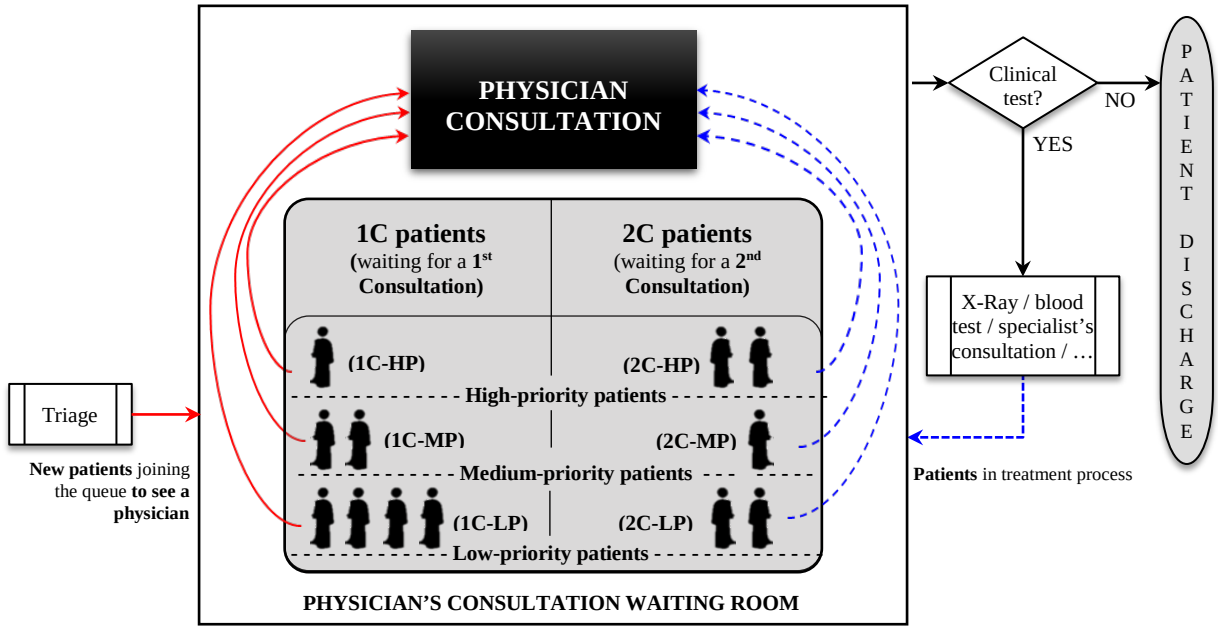


Figure 5.1.Physician consultation queue structure: different priority categories of patients in two different stages.

Pure priority rules

The simplest queue disciplines are those based on pure priority rules. They are also the easiest to implement, which is very convenient in a dynamic and stressful environment such as the ED, especially when physicians have to apply them. A pure priority discipline defines the total order among the categories of patients and chooses the first patient in the non-empty highest priority category. This total order has to be compatible with the partial order induced by the different illness acuity levels in each process stage. That is, in the total order 1C-HP < 1C-MP

< 1C-LP and 2C-HP < 2C-MP < 2C-LP; however, this order can be reversed between different consultations, that is, 1C-MP < 2C-HP could be possible. There are 20 different pure priority disciplines satisfying this partial-ordering condition.

In this study, we consider the four more meaningful pure priority disciplines, named PR-1C, PR-2C, PR-AI, and PR-HN. Table 5.1 contains a full description of the order in which each category of patients is chosen according to each one of these four pure priority disciplines. The 1st consultation pure priority (PR-1C) rule always prioritizes a first consultation over a second one; thus, the order among categories is as follows: 1C-HP, 1C-MP, 1C-LP, 2C-HP, 2C-MP, 2C-LP. The 2nd consultation pure priority (PR-2C) rule always prioritizes the second consultation over the first one; thus, the order among categories is as follows: 2C-HP, 2C-MP, 2C-LP, 1C-HP, 1C-MP, 1C-LP. The acuity index pure priority (PR-AI) rule prioritizes the patients according to their illness acuity index, and within each priority, it prioritizes the 1st consultation over the 2nd consultation; thus, the order among categories is as follows: 1C-HP, 2C-HP, 1C-MP, 2C-MP, 1C-LP, 2C-LP. Finally, PR-HN is the one that is generally followed by the majority of physicians in the HCN, which combines the PR-AI for HP patients with the PR-2C for the MP and LP patients.

Table 5.1. Ordering induced according to the types of patients by several pure priority disciplines.

Discipline	Order induced in the patient categories					
	1st	2nd	3rd	4th	5th	6th
PR-1C	1C-HP	1C-MP	1C-LP	2C-HP	2C-MP	2C-LP
PR-2C	2C-HP	2C-MP	2C-LP	1C-HP	1C-MP	1C-LP
PR-AI	1C-HP	2C-HP	1C-MP	2C-MP	1C-LP	2C-LP
PR-HN	1C-HP	2C-HP	2C-MP	2C-LP	1C-MP	1C-LP

Each one of these priority disciplines is focused on achieving a different objective. Discipline PR-1C attempts to hierarchically minimize the APT by prioritizing all the first consultations. Discipline PR-2C hierarchically minimizes the number of patients in the ED by discharging patients as soon as possible, giving priority to all second consultations to minimize their waiting time in the system. Discipline PR-AI focuses on providing the best possible treatment to the higher priority patients according to the acuity index, assuring the APT limits first and then minimizing the TWT in the ED.

Accumulation priority queues

The APQ management policy generalizes the pure priority queue discipline by setting a discipline based on priority points (PP) that patients of class *i* accumulate at a rate β_i, where $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_k$, and k is the number of different classes of patients. A class-i customer arriving at time t_0 has accumulated $\beta_i(t - t_0)$ PP by time t. When the physician finishes a consultation, the next patient to be seen is the one with the highest PP. Clearly, the APQ model includes the FCFS discipline, obtained by setting $\beta_1 = \beta_2 = \cdots = \beta_k$, and the pure priority disciplines, obtained by setting $\beta_i = M * \beta_{i+1}, i = 1, \ldots, k-1$ and M to a sufficiently large

value. Between both extremes of relationships among the set of beta parameters (equality and very large differences), it is possible to select appropriate values for them to weigh the waiting time, which allows them to overtake a higher priority patient.

In this study, we also propose a modification of the APQ policy that takes into account the time limit targets for each priority. It is motivated by the empirical study of Ding et al. [60], already described in the Introduction section, that analyses the patient routing behaviours of ED decision makers in four EDs using CTAS in Canada. They found that the behaviour of the routing decision makers is best fit by a piece-wise linear concave marginal waiting cost function for each triage level, in which marginal waiting cost has a significantly positive slope below the point where the slope changes and is nearly constant above the CTAS triage-level target wait times. We name this APQ modified policy as the *APQ with finite horizon* policy and denote it with APQ-h. Therefore, the difference between this new policy and the original APQ is that the accumulation of PP at a constant rate finishes at the waiting time limits for the first consultation. From then on, no more priority is accumulated which remains at the maximum value that can be attained in for patients waiting for the first consultation. However, as there is no waiting time target for the second consultation, the limitation of PP does not apply for patients waiting for the second consultation. This truncated APQ model is represented in Figure 5.2. This model helps to stress the capacity of the APQ policy to allow a lower triage-level patient who has waited longer to overtake a higher triage-level patient who has waited less time.

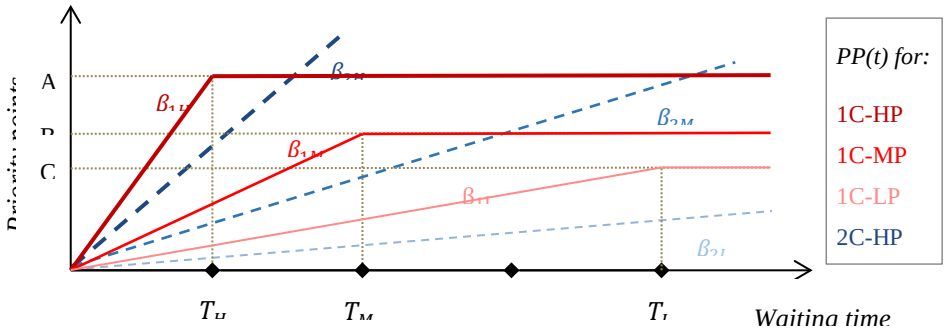


Figure 5.2. Accumulation of priority points with the APQ-h policy for patients classified in three acuity levels.

Therefore, an APQ-h discipline, as well as an APQ discipline, is determined by the vector $\boldsymbol{\beta}$ of the slopes at which the different categories of patients accumulate PP. In our setting with 6 categories of patients, $\boldsymbol{\beta} = (\beta_{1H}, \beta_{1M}, \beta_{1L}, \beta_{2H}, \beta_{2M}, \beta_{2L})$, and in the case of the APQ-h discipline, the parameters associated with the first consultation can be replaced by parameters A, B and C, denoting the maximum PP accumulated at the time limits for the first consultation (see Figure 5.2).

Determination of the optimal APQ-h discipline

In this subsection, we address the problem of finding the optimal values for the vector of parameters $\boldsymbol{\beta}$ that determines the APQ-h discipline with best performance according with the KPIs defined in Subsection 5.3.2. The problem is multi-objective and of a stochastic nature. The necessary notation to define the optimization problem is:

$i \equiv$ index denoting the class of the patient according to the illness acuity index, $i = 1,2,3$ refer to patients of high, medium and low priority, respectively.

$\bar{\lambda}_i \equiv$ average arrival rate of patients of class i.

$T_i \equiv$ APT limit (1st consultation time limit) for patients of priority i.

$\tau_i \equiv$ waiting time for the 1st consultation for a patient of priority i.

$\alpha_i \equiv$ probability that a priority i patient is discharged after the first consultation.

$\nu_i \equiv$ waiting time for the 2nd consultation for a patient of priority i.

$X_i \equiv \begin{cases} 0 \; if \; \tau_i \leq T_i \\ 1 \; if \; \tau_i > T_i \end{cases}$

$E(X_i) \equiv$ ratio of patients of priority i exceeding the APT limit, T_i.

$P_i \equiv$ target for the ratio of patients of priority i exceeding their APT limit.

$E(\tau_i) \equiv$ expected TWT for priority i patients who only need one consultation.

$E(\tau_i + \nu_i) \equiv$ expected TWT for priority i patients needing two consultations with a physician.

$E(TWT_i) = \alpha_i \mathrm{E}(\tau_i) + (1 - \alpha_i)\mathrm{E}(\tau_i + \nu_i) \equiv$ expected TWT for a patient of class i.

$\beta_{1i} \equiv$ slope of the linear accumulating priority function for priority i waiting for the 1st consultation.

$\beta_{2i} \equiv$ slope of the linear accumulating priority function for priority i waiting for the 2nd consultation.

The decision variables of the optimization problem are the slopes β_{1i}, β_{2i}. The time limits T_i and the ratios P_i are the parameters of the problem reflecting the service quality goals, and the expectations $(E(X_i) - P_i)^+$, $E(\tau_i)$, and $E(\tau_i + \nu_i)$ are the functions to be minimized.

Then, the problem of finding the optimal APQ-h (and APQ) management policy, particularized to the case with three types of patients and two consultations, can be formulated as follows in (1):

$$\min_{\boldsymbol{\beta}} \begin{Bmatrix} (E(X_1) - P_1)^+, (E(X_2) - P_2)^+, (E(X_3) - P_3)^+, \mathrm{E}(\tau_1), \mathrm{E}(\tau_2), \mathrm{E}(\tau_3), \\ \mathrm{E}(\tau_1 + \nu_1), \mathrm{E}(\tau_2 + \nu_2), \mathrm{E}(\tau_3 + \nu_3) \end{Bmatrix} \quad (1)$$

We address this multi-objective problem by the weighted sum method, as follows:

Problem [P1]

$$\min_{\boldsymbol{\beta}} W\left(u_1\bar{\lambda}_1\Delta_1 + u_2\bar{\lambda}_2\Delta_2 + u_3\bar{\lambda}_3\Delta_3\right) + \left(v_1\bar{\lambda}_1 E(TWT_1) + v_2\bar{\lambda}_2 E(TWT_2) + v_3\bar{\lambda}_3 E(TWT_3)\right) \quad (2)$$

where $\Delta_i = (E(X_i) - P_i)^+ = \max\{(E(X_i) - P_i), 0\}$.

The weight W expresses the importance of exceeding the goals P_i compared to reducing a time unit of the total waiting time. The sets of weights $\{u_1, u_2, u_3\}$ and $\{v_1, v_2, v_3\}$ indicate the relative importance of achieving each objective in each type of patient. We will consider that the importance of the types of patients is objective-independent and then $u_i = v_i$. Moreover, each patient category is weighted according to their average arrival rate $\bar{\lambda}_i$.

The objective function has no explicit expression in terms of the decision variables. It is a stochastic function that needs to be evaluated by simulation. Therefore, a simulation based optimization (SBO) methodology is used to solve the optimization problem [P1]. SBO is a tool typically used for analysis in the manufacturing context but has not been used often in healthcare system analysis, although it has already been used to find the optimal assignment of resources in EDs (e.g., [32]) and to find optimal management policies for hospital departments, (e.g., [205]–[207], in the case of intensive care units).

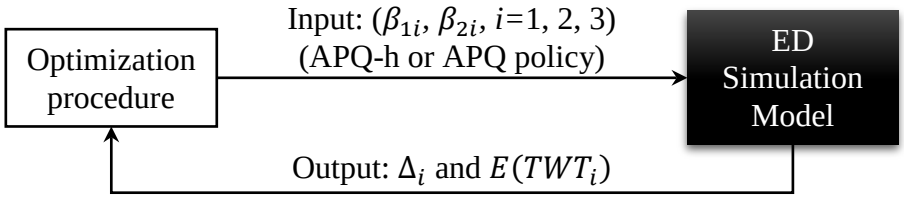


Figure 5.3. Simulation Based Optimization approach.

The rationale of the SBO methodology is as follows: the optimization procedure proposes values for the slopes β_{1i}, β_{2i}, that define an APQ-h (or APQ) policy, which is the input for the simulation model. The ED is simulated under this APQ or APQ-h policy and the outputs – KPIs – are recorded and used to evaluate the random objective. Then, the optimization procedure uses this information and the history of the solutions already evaluated to decide the next solution – APQ or APQ-h policy – to be assessed by the simulation model. This process continues until the stopping conditions of the optimization method are met (see Figure 5.3).

5.4 Case study

5.4.1 Description of the ED

The efficacy of the APQ-h management policies, as well as their comparison with the pure priority rules, are tested by using a simulation model that represents the ED of the HCN. The construction of this simulation model is described in Chapter 2. Moreover, the management policy currently followed by the majority of physicians in the HCN's ED, especially in days of severe overcrowding, does not achieve the goals set by the ED managers, and therefore better management policies should be investigated. As mentioned in Chapter 2, the ED of the HCN - as in many other EDs - it organizes the patient care into two different care circuits: one for the more critical patients, i.e., circuit B, and another for less critical patients, i.e., circuit A. In this chapter's study, we focus on care circuit A, which has its own staff that is not shared with the circuit B and treats patients of priorities 3, 4 and 5.

In the studied care circuit A, there are five exploration rooms and a senior physician in each exploration room**.** The patient routing within the care circuit is the same as that described in Figure 2.8, with patients of priorities 3 (P3), 4 (P4), and 5 (P5) arriving to the ED system, which correspond to the high, medium, and low priorities in Sections 5.2 and 5.3. The next subsection shows the simulation model adaptation for adequately modelling the ED (patients, medical stuff, care paths, etc.) for the described aim.

Data analysis

As mentioned in Chapter 2, the patient arrivals are modelled as a nonhomogeneous Poisson process (NHPP) for each type of patient, with the intensity of arrivals $\lambda_i(t)$ depending on the patient class *i*, and the hour of the day t (there is even a different pattern depending on the day of the week). This seasonality, also observed in other studies depends on the acuity level of the patients. Figure 5.4 shows the arrival rates per hour for the three types of patients of circuit A across the three types of days (holidays, day after a holiday and a normal work day). The average arrival rate of patients across the day (8:00-21:00) and week is 12.17 patients per hour.

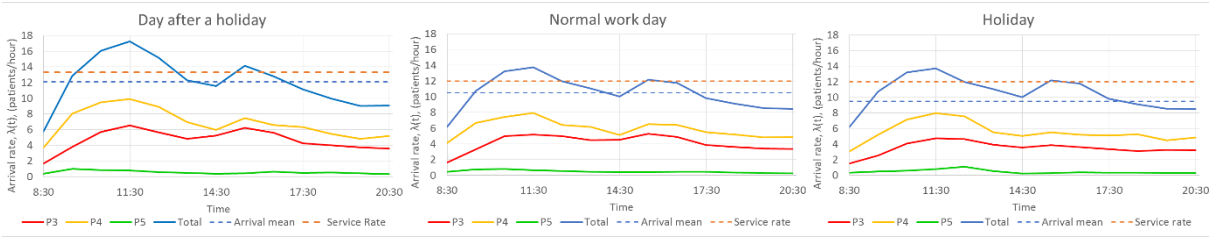


Figure 5.4. Arrival rates of patients, total and according to priority, and service rates for each type of day.

The average service utilization across the day is 90.8%, but the arrival rate is above the service rate for 3 hours (10:00-13:00). The maximum arrival rate peak occurs at the hourly interval

from 11:00-12:00, with a value of 129.87% of the service rate (see Figure 5.4). Table 5.2 contains the quantitative description of the patient flow through the ED, including the probability distributions for first and second consultations service time and the discharge probabilities (α_i) after the first consultation of each priority i patient. Both consultations' service times follow a lognormal distribution with different location parameter value (μ) and the same scale value (σ), which leads to a different expected duration.

Table 5.2. Percentage type of patient (day after a holiday), parameters of the lognormal distribution for the consultation duration and discharge probability after C1.

Priority i	$\%_i$	Service time (min) for the first consultation (S_{1i}): lognormal		Service time (min) for the second consultation (S_{2i}): lognormal		Discharge probability after the 1st consultation (α_i)
		μ_{1i}	σ_{1i}	μ_{2i}	σ_{2i}	
3	38.76	2.89	0.45	2.29	0.45	0.361
4	56.56	2.71	0.45	2.12	0.45	0.513
5	4.67	2.49	0.45	1.89	0.45	0.177

The circuit A service rate of each day - which is calculated from the estimated service time for each patient type and the mix of patients of each type of day - is slightly different from one to another. In this study, we will focus on the most adverse day, the days after a holiday (generally Mondays), in which a service rate of 2.66 patients per hour and physician (13.30 in total, since there are five physicians scheduled all day) is obtained. Moreover all studied KPIs are calculated from 8:00 to 21:00, as during night the ED is not very crowded.

5.4.2 Simulation model

The discrete event simulation (DES) model constructed in Chapter 2 is adapted to assess the performance of the ED under different queue disciplines and under different working and demand pressure conditions. The selection of the next patient to be seen by a physician is simulated by following the rules of the queue discipline that is implemented in the simulation model. Moreover, the flexibility of the simulation model built allows the modification of the mix of patients, seasonality of the arrivals, and the level of congestion. Thus, the robustness of a queue discipline can be investigated by assessing its performance in a wide range of ED scenarios (see Section 5.5).

To avoid the impact of other factors, different from the management of patient rule tested, there are parts of the care process that are simplified and kept out of the limits of the simulation model for his study. For example, the variability of medical test performed outside the limits of the ED, which are not the responsibility of the ED physicians. Thus, in this chapter, the stochastic delay that this additional tests suppose is randomly simulated following a triangular distribution (30, 60, 90) minutes. The simulated priority queue system model is represented in Figure 5.5.

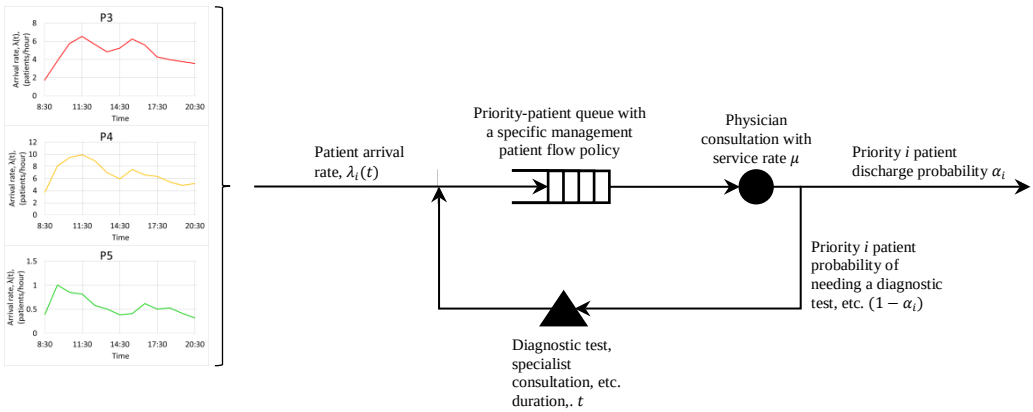


Figure 5.5. ED priority queue model.

The events simulated for this study are as follows:

- *Arrival of a new patient* to the ED with properties such as the priority level and the number of consultations needed. The registration and triage process times are very small, and patients never queue; then, in the simulation model, these times are neglected. Therefore, if any of the physicians are idle at the patient arrival time, the patient enters the first consultation; however, if all physicians are occupied, then the new patient joins the queue in the waiting room.
- *End of a physician consultation.* The patient is then discharged or exits the ED to begin the complementary diagnostic tests. The physician begins a new consultation if there are any patients waiting.
- *Re-entry of a patient to the physicians' waiting room* after medical test are carried out, and the results are ready. At this moment, the patient joins the queue in the waiting room, or the second consultation begins, if there is an available physician.

To determine the simulation run length necessary to accurately estimate the KPI, a preliminary analysis was carried out by running the simulation model for 15,000 days. The KPI estimations were collected and graphically represented as a function of the number of simulated days to identify the stabilization point. As result of this analysis, it was determined that 2,000 simulation days are enough to obtain good KPI estimations (see Figure 5.6).

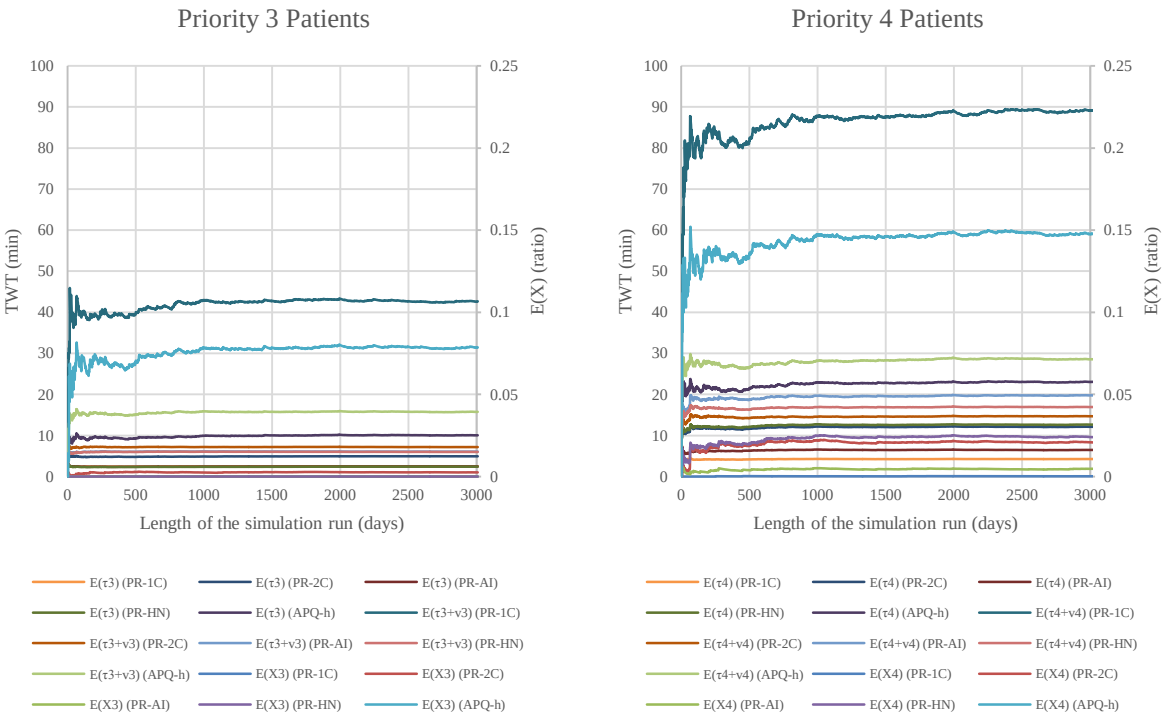


Figure 5.6. Estimation of the KPI values as a function of the number of simulated days.

5.4.3 Optimal prioritization policies

The problem [P1] is to find the optimal APQ-h policy to manage the patient flow of an ED such as the HCN's described in 4.1. In particular, we consider the management of the healthcare circuit of lower priority patients (levels 3, 4, and 5). The values of the parameters included in the objective function are as follows: the average hourly arrival rates $\overline{\lambda_3} = 4.68$, $\overline{\lambda_4} = 6.83$, and $\overline{\lambda_5} = 0.56$; probability of being discharged after the first consultation $\alpha_3 = 0.36$, $\alpha_4 = 0.51$, and $\alpha_5 = 0.18$; maximum APT limits $T_3 = 30$ minutes, $T_4 = 60$ minutes, and $T_5 = 120$ minutes; and the maximum ratio of patients exceeding the APT limits $P_3 = 0.10$, $P_4 = 0.15$, and $P_5 = 0.20$. In addition, the values to weigh the importance between both terms in the objective function and the relative importance of achieving each objective in each type of patient were determined by the ED physician who is a member of the research group q-UPHS, following a discussion with her colleagues. Specifically, the objective independence of the weights for each patient priority was set, that is, $u_i = v_i$. Additionally, the objectives for priority 3 patients were set to be twice as important as for priority 5 patients, and priority 4 patients were 50% more important than priority 5 patients. Therefore, the priority weights were adjusted as follows: $u_3 = 2u_5$, $u_4 = 1.5u_5$, and $u_5 = 1$. Finally, a weight of W=5 was assigned when the patients exceeded the time limit, which is expressed in percentage ($ratio \times 100$). The time unit is expressed in half-hours. Therefore, the increment of 1% in the patients

that exceed the first consultation time limit is equivalent in the objective function to an increment of 2.5 hours in the total waiting time. The instance of [P1] that is solved is as follows:

$$\min_{\boldsymbol{\beta}} 5\,(9.36\,\Delta_3 + 10.245\,\Delta_4 + 0.56\,\Delta_5) + (9.36\,E(TWT_3) + 10.245\,E(TWT_4) + 0.56\,E(TWT_5)) \quad (3)$$

where $\Delta_3 = 100\,\max\{(E(X_3) - 0.10), 0\}$; $\Delta_4 = 100\,\max\{(E(X_4) - 0.15), 0\}$;

$\Delta_5 = 100\,\max\{(E(X_5) - 0.20), 0\}$ and,

$$X_3 \equiv \begin{cases} 0\ if\ \tau_3 \leq 1 \\ 1\ if\ \tau_3 > 1 \end{cases}; \quad X_4 \equiv \begin{cases} 0\ if\ \tau_4 \leq 2 \\ 1\ if\ \tau_4 > 2 \end{cases}; \quad X_5 \equiv \begin{cases} 0\ if\ \tau_5 \leq 4 \\ 1\ if\ \tau_5 > 4 \end{cases}$$

The parameters β are required to sum to 10 to facilitate the comparison of the results among the different scenarios and avoid multiple optimal solutions. The SBO technique described in 3.4 was implemented in the ARENA simulation software ([51], Version 15), which is a suitable software to implement discrete event simulation models and OptQuest optimization software, which is based on the scatter search metaheuristic, as proposed by Laguna and Martí [208].

The optimal values for the maximum priority accumulated by patients waiting for the first consultation are 51.801 (β_{13}=1.7), 33.288 (β_{14} =0.5548) and 16.86 (β_{15} =0.1405), for patients of priority 3, 4 and 5, respectively. Then, the priority accumulated by a priority 3 patient in almost 20 minutes equals the priority accumulated by a priority 4 patient in 60 minutes and by a priority 5 patient in almost three hours. The slopes for the second consultation are 7.5775, 0.0005 and 0, which means that priority 5 patients are only seen by the physician for a re-evaluation when the ED no longer has higher priority patients. Therefore, the relative importance that should be given to the different stages of the care process is not the same for all priorities, that is, the dominance relation between β_{1i} and β_{2i} is not always the same ($\beta_{13} < \beta_{23}$ while $\beta_{14} > \beta_{24}$).

The optimization process was also applied to determine the optimal value of the APQ parameters, but no significant difference was found in the ED performance (KPI values) when using the optimal APQ-h. Thus, there is no practical difference in applying any of both queue disciplines.

The simulation results for the KPI for pure priority disciplines and the optimal APQ and APQ-h disciplines are shown in Table 5.3. Disciplines APQ-h and PR-1C are able to achieve the goals for the probability of patients exceeding the time limit but they are not achieved by the other disciplines, including the currently used one. It should be noted that particularly, the currently used pure priority rule in the HCN's ED, PR-HN, has their KPIs out of control and are considerably improved with the optimized new policy APQ-h (and also by the APQ).

Table 5.3. KPI for pure priority disciplines and APQ and APQ-h.

Discipline			PR-AI	PR-1C	PR-2C	PR-HN	APQ & APQ-h
Ratio of patients exceeding the time limit P_i. ($P_3 = 0.1, P_4 = 0.15, P_5 = 0.2$)		$E(X_3)$	<0.001	<0.001	0.004	<0.001	0.061
		$E(X_4)$	0.111	0.016	**0.300**	**0.306**	0.148
		$E(X_5)$	**0.457**	0.066	**0.452**	**0.453**	0.199
Total waiting time	patients who need a single consultation	$E(\tau_3)$	2.838	2.885	5.458	2.747	9.552
		$E(\tau_4)$	21.474	9.683	44.564	45.420	26.494
		$E(\tau_5)$	171.137	28.978	166.449	168.034	54.699
	patients who need medical tests (2 consultations)	$E(\tau_3 + \nu_3)$	7.596	65.038	8.021	7.394	18.816
		$E(\tau_4 + \nu_4)$	129.149	155.640	47.341	51.224	113.371
		$E(\tau_5 + \nu_5)$	235.253	226.350	169.732	172.291	223.557
Objective function value			103.723	44.511	858.719	879.589	31.990

The APQ and APQ-h policies' KPIs $E(X_i), i \in \{3,4,5\}$ fall within the limits; however, as is expected, the PR-1C is the only pure priority policy whose KPIs $E(X_i)$ are also within the boundaries. This policy focuses on assisting patients waiting for the first consultation, which leads to better values for $E(X_i)$ –further from the limits – while the rest of the performance KPIs and, consequently, the objective function value are significantly worse.

Display and interpretation of the simulation results using star graphs

In this subsection, the ED performance when using pure disciplines to manage the patient flow of an ED such as the HCN is compared with the ED performance when using the optimal APQ-h, combining the three factors taken into account. The simulation results for the KPIs and the value of the objective function are displayed by using a star graphs (see Figure 5.7). The upper vertical axis (*OFV*) represents the objective function value while the first three axes clockwise show the ratios of patients exceeding the APT limit for each priority ($E(X_3)$, $E(X_4)$, and $E(X_5)$). The values below the upper limits for each priority objective are in black and those above them are in red.

The next three axes clockwise ($E(\tau_3), E(\tau_4)$, and $E(\tau_5)$) represent the expected TWT in the ED system of patients who only need a single consultation with the physicians, while the last three axis ($E(\tau_3 + \nu_3), E(\tau_4 + \nu_4)$, and $E(\tau_5 + \nu_5)$) display the expected TWT in the ED system of patients who are not discharged after the first consultation. Because the goal is to minimize all KPIs and the objective function, the nearer each value is to the centre of the chart, the better the performance is.

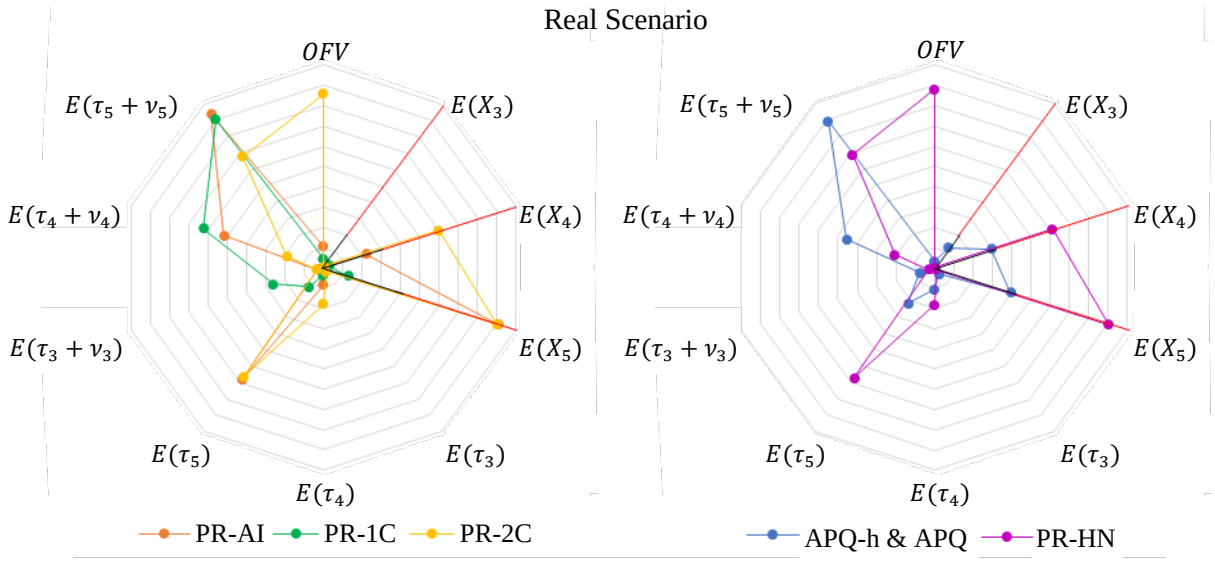


Figure 5.7. Star plot of the simulated results of the real ED scenario of the HCN.

Figure 5.7 displays and compares the KPIs obtained by the analysed queue disciplines in the real HCN scenario. The star plot on the right displays the KPIs for the PR-HN (the rule which is used by the majority of the medical stuff in the HCN) and the optimal APQ-h and APQ, while the star plot on the left displays those for the pure priority rules PR-1C, PR-2C, and PR-AI. The APQ-h policy provides results for $E[X_i]$ that are lower than but close to the P_i boundaries, which produces both no penalties and room to improve the results in the other KPIs. The discipline PR-1C respects the P_i limits, while the PR-AI, PR-2C, and PR-HN policies do not (PR-AI: $E(X_5) = 0.457 > 0.2$; PR-2C: $E(X_4) = 0.300 > 0.15, E(X_5) = 0.452 > 0.2$; PR-HN: $E(X_4) = 0.306 > 0.15, E(X_5) = 0.453 > 0.2$).

However, because PR-1C prioritizes the first consultation, the $E[\tau_i + \nu_i]$ values are worse than those obtained by APQ-h and APQ (65.04 vs 18.82, 155.64 vs 113.37, and 226.35 vs 223.557 for $E[\tau_3 + \nu_3], E[\tau_4 + \nu_4]$, and $E[\tau_5 + \nu_5]$), respectively. As a consequence, the APQ-h policy obtains a better global performance, as measured by the value of the objective function.

5.4.4 Sensitivity analysis for the criteria's importance

In this section, we analyse the robustness of the optimal solutions to the APQ-h parameters when the weight W in (2) is varied. The weight fixed by the physicians (W=5) – whose main objective is to achieve the performance level for the APT limits – is considered to be one of the extreme values for the range of studied values. From that point, the weight is being reduced until a minimum of 0.2 is reached (at this point the worsening of 1% in the number of patients that exceed the time limit for first consultation is equivalent to increase the total waiting time in 6 minutes).

W is varied in the range [0.2, 5] and the following two different optimal solutions are found: one is optimal for W varying from 0.2 to 0.3 and the other is optimal for W varying from 0.4 to 5. The solution for $W \geq 0.4$, provides $\Delta_i = 0, i = 3, 4, 5$, and then the first part of the objective is fully minimized with a value of zero. The solution is that mentioned in the previous section (the optimal solution for the case study). Figure 5.8 shows the KPI values for each solution trough the interval of values [0.2, 0.7] with steps of 0.1 (note that there is no change from 0.4 onwards).

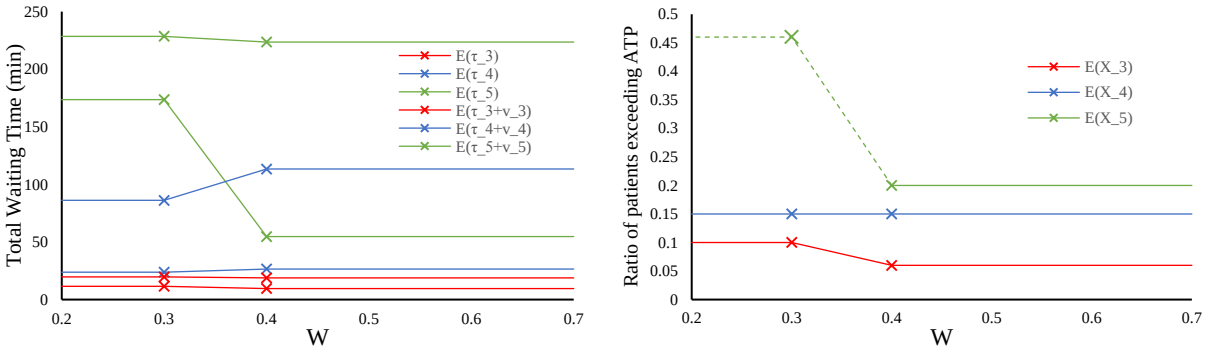


Figure 5.8. Outcomes of the optimal solution for different objective functions (W ranging from 0.2 to 0.7 as there is no change from 0.4 onwards): total waiting time in the system (left graph) and ratio of patients exceeding the time limit for the first consultation (right graph). The latter represents the values that does not achieve the target for the ratio (above the limit) in dashed line and those that does (equal or below the limit) in solid line. The crosses indicates the change-points.

However, from W =0.2 to W =0.3, $\Delta_5 > 0$ due to the domination of the objective function by the minimization of the TWT (objective 2). The TWT of priority 4 patients is significantly reduced ($E(\tau_4)$ is reduced from 26.494 to 23.776 and $E(\tau_4 + v_4)$ from 113.371 to 86.179), which represents 57% of the patients. The priority 3 patients' waiting time is almost the same, while the waiting time of priority 5 patients worsens. The optimal solution in this case is $\beta_{13} = 1.359$, $\beta_{14} = 0.6580$, $\beta_{15} = 0$, $\beta_{23} = 8.0051$, $\beta_{24} = 0.0010$, and $\beta_{15} = 0$. Contrary to the previous solution, in this case, priority 4 patients who are waiting for their second consultation have a greater accumulating priority rate than all priority 5 patients, who are only assisted if there are no other patients in the ED.

5.5 Extended simulation study to a general set of ED scenarios

5.5.1 Selection of scenarios

In this section, an extended analysis of the performance and a comparison between pure priority disciplines and optimal APQ-h disciplines are carried out in different ED scenarios. The set of scenarios is designed from the HCN ED, which is described in Section 5.4, by varying the

average occupancy rate, the pattern of the intraday seasonality, and the composition of the mix of patients. Specifically, we consider the following values for the abovementioned factors:

- **ED congestion level** (named as factor 1 and denoted by **F1**): The average occupation rates ρ of 90% and 95% are considered. The number of physicians is maintained while the patient arrival rate is modified accordingly.
- **Arrival seasonality (F2):** Three arrival seasonality patterns are considered, ranging from no seasonality (constant arrival rate of patients, denoted as *T0*) to a maximum hourly seasonality, which is described by two different triangular patterns for the arrival rate $\lambda(t)$, both with a peak at 11:30 a.m. and a ratio of $(\lambda_{max} - \lambda_{min})/\lambda_{min} = 0.5$. The first triangular pattern, denoted as Tu, extends the triangular shape across the entire time range, while the second triangular patterns, denoted as Tp, only applies the triangular shape in the time range [10:00, 13:00], with the arrival rate out of this range being constant (see Figure 5.9). As a consequence, each one of the three seasonality patterns have different values for λ_{max}.
- **Mix of patients (F3):** Four different mixes of patients are considered as follows: balanced distribution among all types of patients (1/3 of P3, 1/3 of P4 and 1/3 of P5) and a biased mix towards each priority (50% of P3, 25% of P4 and 25% of P5; 25% of P3, 50% of P4 and 25% of P5; and 25% of P3, 25% of P4 and 50% of P5). These scenarios are denoted by B0, B3, B4 and B5, respectively.

Each scenario is denoted by a vector (f_1, f_2, f_3), where f_i is the level of factor $F_i, i = 1,2,3$, and $f_1 \in \{90\%, 95\%\}$, $f_2 \in \{T0, Tu, Tp\}$, and $f_3 \in \{B0, B3, B4, B5\}$. Then, a total of 24 scenarios will be analysed with the simulation model.

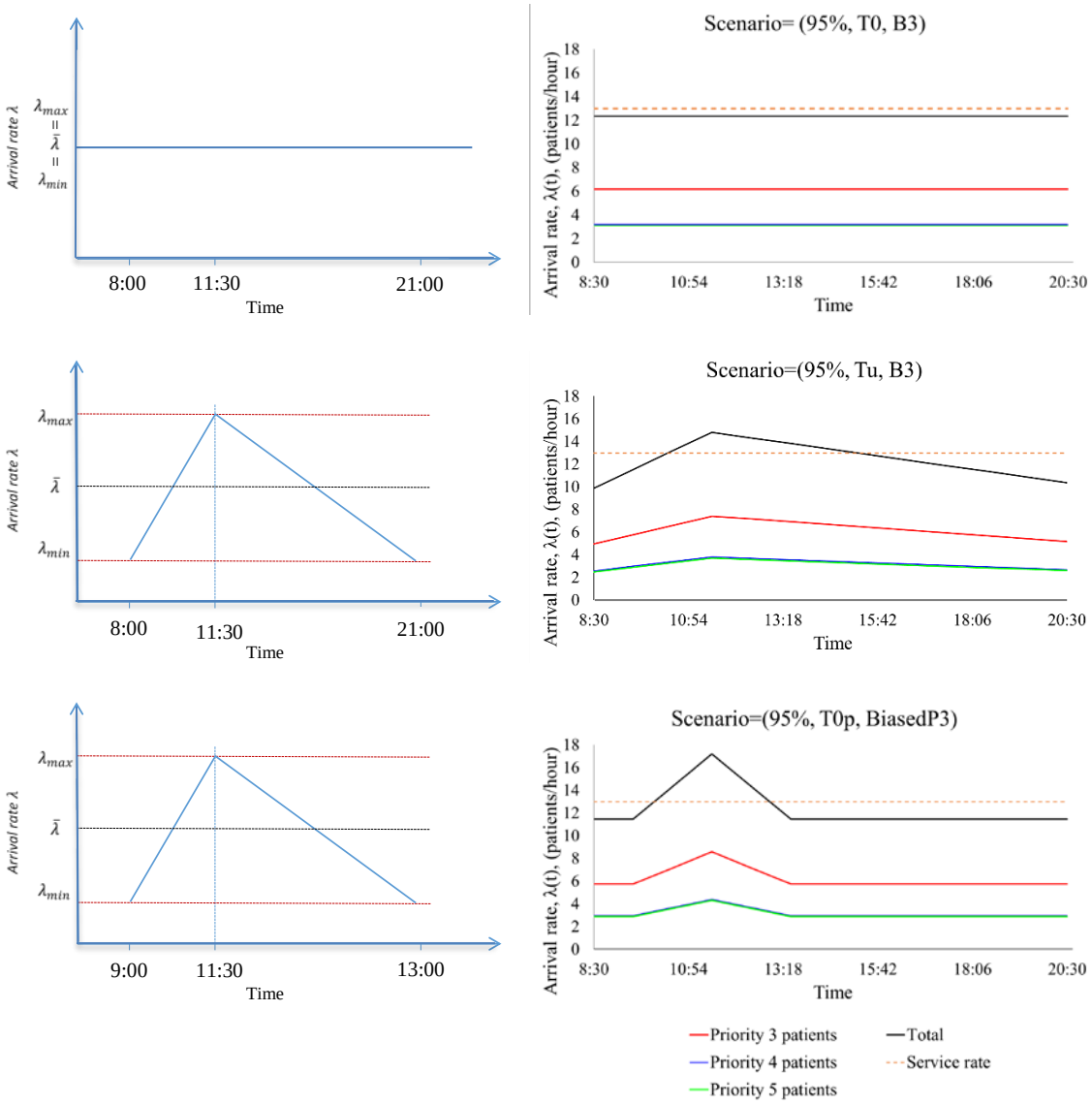


Figure 5.9. The three patterns for the seasonality of the arrivals: from top to down scenarios T0, Tu and Tp, respectively.

5.5.2 Analysis of the results

Influence of the demand factors on the ED performance. The results from the simulation of the different ED scenarios show the influence of the demand factors (quantity, seasonality and typology) on the ED performance. This influence is visualized in Figure 5.10, which displays, by using star graphs, the KPIs and the objective function of a selected set of ED scenarios. The scenarios in the first row are of type (90%, Tu, -); that is, they differ in the mix of patients. The scenarios in the second row are of type (95%, Tu, -); that is, they only differ from the scenarios in the first row in the congestion level, that is, of 95%. Finally, the three scenarios in the third row are of type (95%, -, B4); that is, they differ in the seasonality pattern for the arrivals. The

following observations can be extracted from this figure and, in general, from all scenarios results:

- The results are very sensitive to the increase in the occupancy ratio from 90% to 95% (the KPIs worsen from the first row, 90%, to the second row, 95%).
- The mix of patients also has a large influence; the higher the severity of patients who represent the maximum percentage in the mix of patients, the worse the performance is (as observed in the first and second rows).
- The seasonality also influences the performance. The best results are observed in the case of homogeneous arrivals, and the worst results are observed in the case of a triangular pattern extended throughout the day.

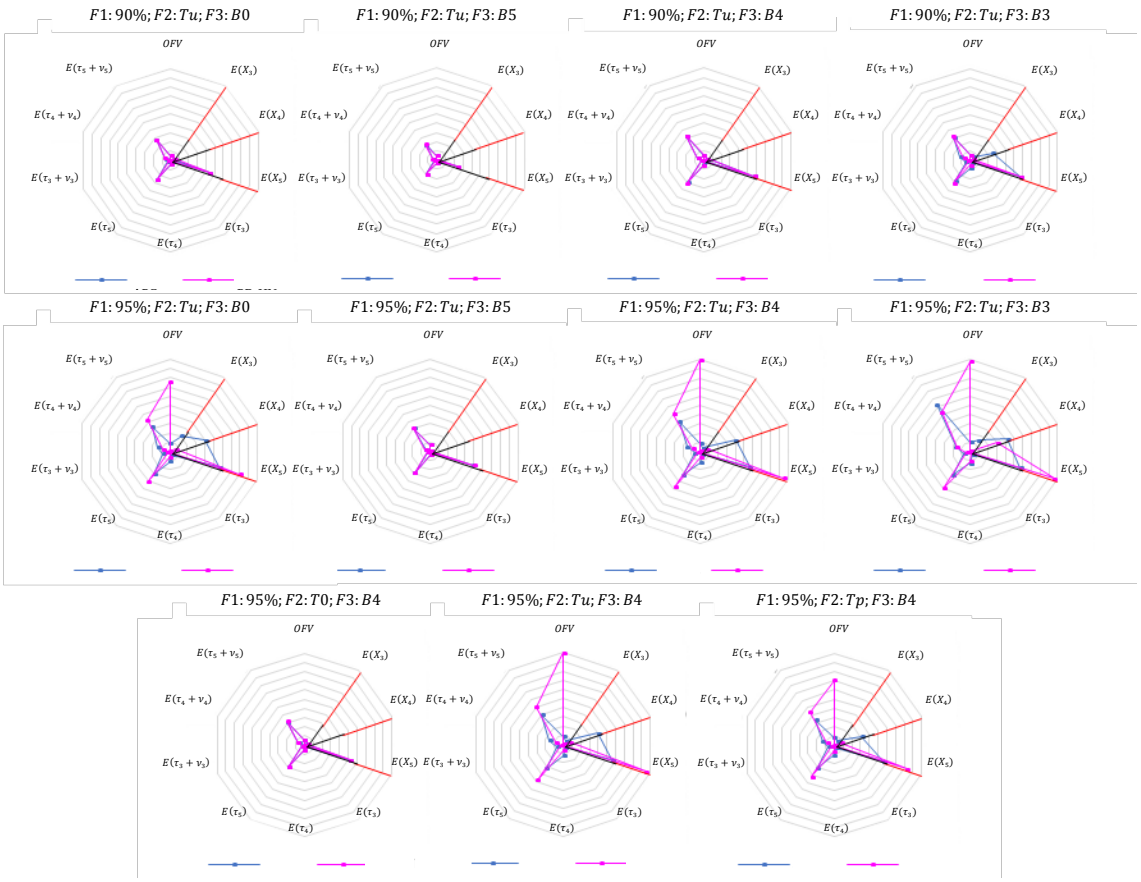


Figure 5.10. Representation of the KPIs in selected scenarios.

APQ-h versus the pure priority disciplines: performance comparison

The simulations of the different ED scenarios ruled with pure priorities and APQ-h disciplines produced results that highlight the very different behaviours of all of them; while each pure priority focused on the achievement of a specific subset of KPIs, disregarding the others, the APQ-h policy is able to balance all the KPIs according with their relative importance expressed through the weights in the objective function. This general statement is graphically visualized

in Figure 5.11, in which the simulated results of all the ED scenarios are represented in the same star plot for each queue discipline.

*The PR-2C policy prioritizes the minimization of the TWT for patients who need medical diagnostic tests***.** The shape created in Figure 5.11 by the KPIs associated with the PR-2C policies is graphically shifted to the right and down, as this pure discipline focuses on discharging patients waiting for the second consultation, that is, on $E(\tau_i + \nu_i)$. In cases of a high congestion and a high percentage of high priority patients, disregarding the first consultation produces the non-fulfilment of the time limits for the APT and, therefore, the results in positive values for the ratio Δ_i.

*The PR-1C policy prioritizes the minimization of the TWT for patients who need a single consultation***.** Opposite to PR-2C, the shape created in Figure 5.11 by the KPIs associated with the PR-2C policies is graphically shifted to the top left, as this pure discipline only pays attention to the APT limit target, ignoring the waiting time for the second consultation, ν_i. Therefore, $\Delta_i = 0$ and the $E(\tau_i)$ values are small but the $E(\tau_i + \nu_i)$ values are large.

The PR-AI policy prioritizes the minimization of the APT and TWT for the highest priority patients. The shape created in Figure 5.11 by the KPIs associated with the PR-AI policies is a mixture of the previous ones, i.e., the best results in all KPIs for the highest priority patients and the worst for the lowest priority patients.

The PR-HN policy prioritizes the minimization of the APT and TWT for the highest priority patients and the minimization of the TWT for patients of other priorities who need medical diagnostic tests. The shape created in Figure 5.11 by the KPIs associated with the PR-HN policies is similar to the shape created by the PR-2C policies. The difference is that the values for the highest priority KPIs, $E(\tau_3 + \nu_3)$ and $E(\tau_3)$, are nearer to the centre of the chart.

*The optimal APQ-h policies produce balanced results***.** The optimal APQ-h policies obtain worse results than PR-1C for $E(\tau_i)$, although they achieve $\Delta_i = 0$ and better results for $E(\tau_i + \nu_i)$. The opposite is concluded when compared with PR-2C, i.e., the results are better for $E(\tau_i)$ and Δ_i but worse for $E(\tau_i + \nu_i)$. When compared with PR-AI, the results are worse for the highest priority patients but better for the lower priority patients. Therefore, the shapes of the star plots associated with the APQ-h policies are more centred and close to the central point than the shapes of the other policies, meaning more balanced results are obtained.

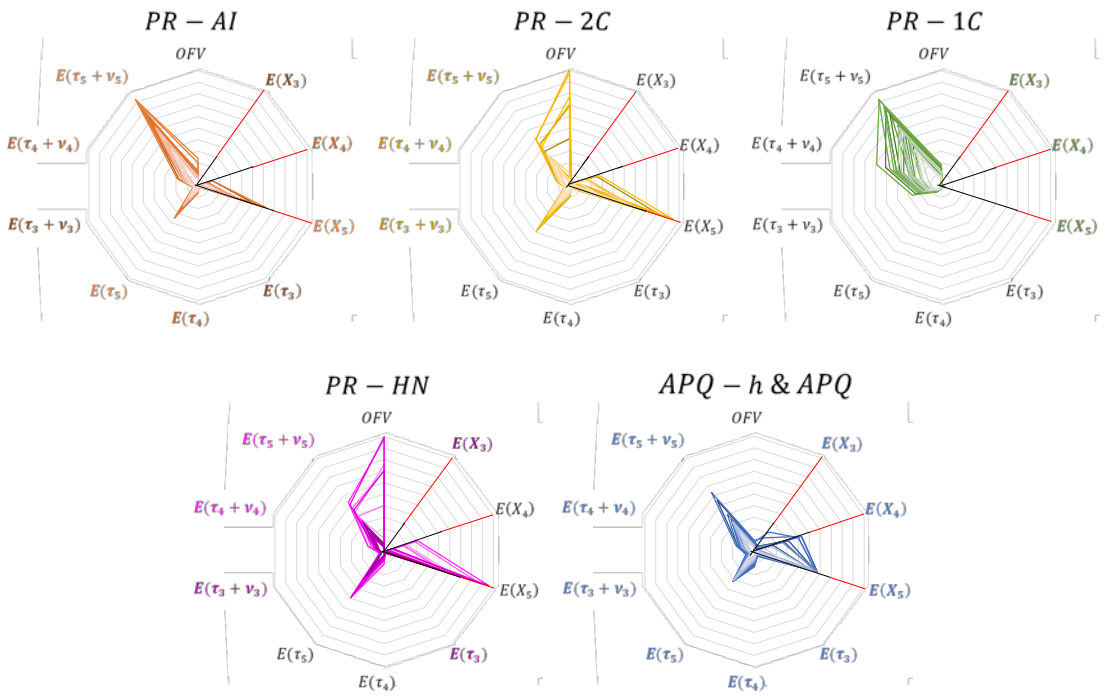


Figure 5.11. Representation of the scenarios KPIs ruled by the PR-2C, PR-AI, PR-1C, PR-HN and APQ-h and APQ policies.

Table 5.4 compares the values of the objective function obtained for each discipline in each of the 24 ED scenarios analysed and quantifies the improvement obtained by the APQ-h policy with respect to the best pure priority policy. Table 9.1 of Appendix H displays the detailed results for each scenario by disclosing the value for every KPI considered. The last columns show the optimal APQ-h policy in terms of the values of β_i. Generally, the highest improvements are obtained in the toughest environments for the ED, that is, the 95% occupation level, seasonality in the patient arrivals and high percentages of high severity patients. For low occupation levels (90%), almost no improvement is achieved, as the optimal APQ-h policy is very similar to the best pure priority rule solution (PR-2C) – giving more importance to the second consultation than to the first one, as it is easy to obtain the time target limit. For example, in the first row scenario, the slopes are 0.0396, 0.0001, and 0 for first consultations (β_{13}, β_{14}, and β_{15} respectively), and 5.7720, 4.1227, and 0.0656 for second consultations (β_{23}, β_{24}, and β_{25} respectively), and the objective function value obtained for the APQ-h policy is almost the same as for the PR-2C policy.

Table 5.4. Summary of the objective value of each scenario with the different queue disciplines and the improvement of the optimal APQ-h with respect to the best pure priority rule.

Scenario			Management Policy						APQ-h solution					
F1	F2	F3	PR-AI	PR-1C	PR-2C	PR-HN	APQ-h	Improvement	β_{13}	β_{23}	β_{14}	β_{24}	β_{15}	β_{25}
0.9	T0	B0	106	145	75	81	75	0%	0.0396	5.7720	0.0001	4.1227	0.0000	0.0656
		B5	113	137	71	75	71	0%	0.0368	4.9629	0.0001	4.9437	0.0000	0.0565
		B4	103	145	76	81	76	0%	0.0443	4.9691	0.0001	4.9539	0.0000	0.0326
		B3	108	161	82	88	82	0%	0.0601	7.8318	0.0001	2.0637	0.0000	0.0443
	Tp	B0	143	205	98	106	98	0%	0.0608	5.7806	0.0001	4.1199	0.0000	0.0386
		B5	158	196	94	99	94	<1%	0.0376	5.3776	0.0001	4.5459	0.0000	0.0388
		B4	139	208	101	108	101	0%	0.0608	4.9572	0.0001	4.9453	0.0000	0.0366
		B3	144	230	108	115	108	0%	0.0601	8.8210	0.0001	1.0723	0.0000	0.0465
	Tu	B0	161	231	107	116	108	0%	0.0602	5.3336	0.0001	4.5612	0.0000	0.0449
		B5	179	220	103	109	103	<1%	0.0407	5.7440	0.0001	4.1624	0.0000	0.0528
		B4	155	236	110	125	111	0%	3.4125	3.9649	0.0013	2.1180	0.0000	0.5033
		B3	158	257	245	252	126	20%	0.1202	9.5409	0.0967	0.1213	0.0001	0.1208
0.95	T0	B0	171	240	114	114	114	0%	0.0479	5.7642	0.0001	4.1353	0.0000	0.0525
		B5	188	227	108	107	108	0%	0.0456	5.9277	0.0001	3.9886	0.0000	0.0380
		B4	164	244	117	117	117	0%	0.0653	9.7665	0.0001	0.0999	0.0000	0.0682
		B3	163	254	119	122	120	0%	0.0603	9.8036	0.0001	0.0681	0.0000	0.0679
	Tp	B0	225	330	895	879	159	29%	0.1211	9.4163	0.0825	0.2582	0.0001	0.1218
		B5	253	315	142	141	142	0%	0.0433	4.9817	0.0001	4.9343	0.0000	0.0406
		B4	217	336	1545	1552	172	21%	0.1217	5.0678	0.0726	4.6173	0.0001	0.1205
		B3	212	349	1522	1538	185	13%	1.3123	5.8177	0.4861	2.3816	0.0007	0.0016
	Tu	B0	248	370	1693	1668	180	27%	0.2402	9.1343	0.1274	0.2565	0.0002	0.2414
		B5	282	352	155	154	155	0%	0.0378	5.3497	0.0001	4.5821	0.0000	0.0303
		B4	435	378	2220	2197	196	48%	0.2401	7.0783	0.126	2.3531	0.0002	0.2023
		B3	522	395	2144	2153	213	46%	1.7532	5.1798	0.1021	2.9624	0.0024	0.0001

The similarity between the optimal APQ-h policy and one pure priority policy is observed in other ED scenarios. However, having no strict priority imposed by the APQ-h policy can enormously affect the KPI results. This fact is illustrated in Table 5.5, which contains the KPI results and APQ-h optimal solution for the (95%, Tu, B0) ED scenario. The optimal APQ-h policy favours the ED discharge of patients by assigning larger slopes to patients waiting for the second consultation, i.e., ((β_{23}=9.1343)>β($_{24}$=0.2565)>(β_{25}=0.2414)>(β_{13}=0.2402)> (β_{14}=0.1274)> β_{15}=0.0002)). However, even if the order of attending patients seems to be similar to the PR-2C discipline, the flexibility of APQ-h allows P3 and P4 patients who have been waiting for their first consultation for a long time to overtake patients who have been waiting for less time for their second consultation. As a result, the PR-2C policy does not fulfil $\Delta_i = 0$, but the APQ-h policy does, and the objective function value is 1693.08 for PR-2C and 180.28 for the APQ-h policy.

Table 5.5. KPI results and APQ-h optimal solution for the (96%, Tu, B0) ED scenario.

Scenario	Queue Discipline	Obj	$E(X_3)$	$E(X_4)$	$E(X_5)$	$E(\tau_3)$	$E(\tau_4)$	$E(\tau_5)$	$E(\tau_3 + v_3)$	$E(\tau_4 + v_4)$	$E(\tau_5 + v_5)$	**Slopes** obtained by solving the optimization problem
F1: 95%; F2:Tu; F3: B0	PR-AI	247.60	0.00	<0.01	0.13	2.46	6.44	45.17	6.08	19.69	169.50	β_{13}=0.2402 **β_{23}=9.1343** β_{14}=0.1274 **β_{24}=0.2565** β_{15}=0.0002 **β_{25}=0.2414**
	PR-1C	370.02	0.00	<0.01	<0.01	2.51	4.28	11.19	41.67	86.75	169.89	
	PR-2C	1693.08	<0.01	0.02	**0.27**	4.94	12.10	78.88	7.24	14.77	82.25	
	PR-HN	1668.13	0.00	0.02	**0.28**	2.40	12.59	79.66	6.04	17.01	85.12	
	APQ-h	180.28	0.07	0.14	0.20	9.96	22.49	60.53	14.64	28.44	66.53	

5.6 Conclusion

In this chapter, the performance of the APQ-h policy in a real ED setting that considers the stochasticity in the arrivals of patients and the different stages of the health care process has been investigated. Therefore, this study extends the theoretical results obtained in studies ([178], [209]) that assumed a homogeneous Poisson process for the arrivals and only one stage for the patient treatment. The results show that ED performance is better when it is managed with the APQ or APQ-h policy than with other priority policies. This observation supports the use of any of both policies in practice to manage the ED patient flow, in fact the APQ-h is already followed in some EDs, as reported in [60]. We identify that managers of the ED hospitals included in Ding et al.'s study [60] apply a structure for queue discipline that is equal to the APQ-h. This policy that might seem counterintuitive because the patients stop accumulating priority points, in fact, states that from that moment on the patients waiting for first consultation are selected by priority and a FIFO rule within each category, which is a rule widely used to manage EDs as-mentioned in the introduction (see, for example, Taylor et al. [180], Haussman [181], Siddharthan and Jones [182], Laskowski et al. [183], Mokaddis et al. [184]). Any patient of high priority having reached the time limit for the first consultation will never be overtaken by other patient of lower priority waiting for the first consultation, independently of their respective waiting times. Only patients having waited for a very long time for their second consultation could overtake such high priority patient. We use a simulation-based optimization methodology to obtain the optimal APQ-h and APQ policies that are superior compared to other pure priorities disciplines, especially when high congestion and non-stationary ED environments are considered. Moreover, in the case study of the ED of the HCN, the use of APQ-h and APQ significantly outperforms the current priority rule, PR-HN, whose obtained KPIs were out of control.

However, the analysis also shows that in not very congested ED scenarios, with a time-regular affluence of patients, the application of APQ-h or APQ has no advantage over the best pure priority policy. In these cases, it is recommended to apply the pure priority discipline because it is easier to implement and is very convenient in a dynamic and stressful environment such as the ED, especially when physicians have to apply them. Furthermore, pure priority disciplines require less information, only requiring the type of patient and the stage of healthcare process but not the recording of the waiting time for each patient, as it is necessary with the APQ-h and APQ policies.

The analysis of the ED performance was carried out by considering several KPIs related to the APT and the waiting time for consultations. Other specific KPIs could be considered to assess the performance of the ED under different patient flow management policies. Nevertheless, the application of pure priority rules is goal- and objective-independent, and therefore, ED performance will remain unchanged. The computational analysis carried out in this chapter shows that these rules can be optimal for certain objectives but provide very bad values for

others. For example, in non-congested EDs, the PR-2C policy works better than the rest of the pure priority rules, while in very congested EDs, the PR-1C policy works better, especially when great importance is given to avoiding exceeding the APT limit. However, by definition, optimal APQ-h and APQ policies are objective- and goal-dependent, which provides them with a flexibility that managers can use to adapt them to achieve specific objectives or to obtain solutions that balance all of them.

The introduced APQ-h discipline is a modification of the APQ discipline, justified by the previous empirical study [60]. In our computational study, we optimized the parameters of both types of policies to compare them and to determine which one is better or in which ED scenarios one outperforms the other, but in all tested scenarios both policies produced the same results. Differences in KPI values were in decimals, which is attributable to the non-exact optimization procedure and the evaluation of the KPIs by simulation. Therefore, given that no practical differences between them have been found in any of the analysed scenarios with the considered objective functions and both of them have been found to be superior to the other pure priority policies, any of both modalities of the APQ policies could be recommended to be implemented for the management of the ED patient flow. However, although we have not found scenarios in which they differ, their structure is somewhat different, and it is possible that in some situations or under different objective functions, one of the two disciplines will surpass the other. This issue remains to be investigated. Moreover, it would also make sense to investigate a non-linear rate for accumulating priority by taking into account the slack of a patient until the APT is exceeded. Finally, the problem has been solved using commercial optimization software, which, in some scenarios, has shown a slow convergence to the *optimal* solution. Therefore, treating the problem from a multi-objective point of view and developing an efficient optimization algorithm to estimate the Pareto frontier remains an objective for future research.

II. PHYSICIAN SCHEDULING PROBLEM

Chapter 6 Scheduling problem definition

6.1 Introduction and related literature

The Emergency Room (ER) of a hospital is where medical and/or surgical care is given to patients arriving in need of immediate attention. An ER is therefore a 24/7 service. Physicians are required to work night, day and weekend shifts, and take on different ER assignments. Complex constraints add to the difficulty of finding good and equitable schedules for the physicians. Examples of ergonomic constraints are described in [210], while [5] offer an overview of other typical constraints to classifying them into four categories: 1) supply and demand, 2) workload, 3) fairness and 4) ergonomics, based on five case studies performed in Canadian hospitals. This part of the thesis addresses a real physician scheduling problem in which constraints of all four categories are considered.

Although the physician scheduling problem shares many characteristics with the nurse scheduling problem (and other workforce planning problems, see, for example, [211], [212]), it has received much less attention in the literature. A review of the nurse rostering problem can be found in [213], [214]. One can, of course, expect the type of techniques that work well in one problem to do just as well in another, but, despite their basic similarity, they also have differences that can condition the solution. A thorough analysis of such differences is provided in [215], which highlights the importance of modeling preferences and fairness, among other issues. Their conclusion is that, its combined characteristics make the physician scheduling problem highly unique, and thus distinct from general personnel scheduling problems.

The planning horizon considered in most published studies tends to be small, ranging from two to four weeks. In [216], for example, the physicians in an anesthesia department are scheduled to cover a two-week planning horizon, later extended to six weeks in [217]. However, our study addresses a one-year planning horizon, that is, much longer than the usual scheduling periods reported in the literature. Twelve-month work calendars are a legal requirement in some countries, including Spain, where they are drawn up annually by the company (after consultation and a subsequent report to the workers' representatives) and available for all to see in the workplace (Article 36 of Workers’ Statute, BOE-A-2015-11430, [218]). This calendar must contain both the work schedule and annual distribution of working days and holidays. It will take into account the maximum number of legal working days, which is determined by collective agreement or work contract. This calendar may undergo modification throughout the year due to changes affecting the staff, family care leave, sickness, etc. In such cases, the

manager has to meet staffing demand with minimum change to the original calendar. However, the operational management of the work calendar is a different problem and beyond the scope of this thesis.

The major solution approaches for solving the physician scheduling problem involve mathematical programming, metaheuristics, constraint programming, and column generation (reviewed in [5]). Similar results are presented by [7], who analyze the characteristics of the problem and scheduling techniques based on Linear Programming (LP) and metaheuristics (mainly Tabu Search). See [215] for a recent review of 68 relevant papers addressing different types of physician scheduling problem in hospitals. They are classified as either Staffing, Rostering, or Re-planning problems. The majority, 61 papers, use mathematical programming models. They can be exactly solved for small instances or for problems that are not heavily constrained. In [219], resident physicians in a hospital's pulmonary unit are scheduled for a 6-month period. The author considers 29 instances with the number of variables ranging from 486 to 1,995 and the number of constraints ranging from 552 to 2,907. Most of the instances are exactly solved within seconds by commercial solvers, but not in all cases, even when the problems are small in size. Nevertheless, the solutions are of high quality and comparable with manually-created schedules, and therefore valid for practical purposes. In other cases, as in [220], where the Integer Linear Programming (ILP) model could not be solved by a modified version of the branch and bound method, due to the large dimension of the problem, a heuristic approach based on a partial branch and bound was used. In fact, when the problem at hand is large (a large number of physicians to be scheduled over a long planning horizon) and very detailed models are formulated, exact solution approaches are usually impractical.

The physician scheduling problem falls into the category of NP-hard problems, which are intractable for large instances. In these cases, a solution can be obtained by heuristic algorithms, usually guided by metaheuristics, or a combination of heuristics and exact methods (see [221]). In a review of methods and models for solving staff scheduling and rostering problems, [222] identified 28 different methods, ranging from integer programming to all types of metaheuristic algorithms. For example, [223] solved the physician rostering problem by using a genetic algorithm for a one-month planning horizon and a small/medium size ER with 16 physicians. [224] models the staff scheduling problem at an Emergency Medical Service using ILP, which is solved by a Variable Neighborhood Decomposition Search heuristic. The results show how the heuristic approach outperforms a state-of-the-art commercial ILP solver. In [225], a simple heuristic is used to assign guard shifts over a one-year horizon. The problem is not heavily constrained and the number of shifts assigned per day is small: two or three depending on the day type.

The physician scheduling problem addressed in this and the following chapters is complex because of the long planning horizon and complicated constraints. In addition to demand constraints, it considers all compulsory constraints imposed by legislation and personnel preferences. The objective is to achieve the fairest feasible schedule. The problem is initially

modeled as an ILP problem, but, after a real instance of this problem remains unsolved by a well-known ILP solver in one week, using a powerful computer, a hybrid algorithm is designed. The constructive phase of a Greedy Randomized Adaptive Search Procedure (GRASP) is designed to obtain full schedules, which are subsequently improved by means of a Variable Neighborhood Descent Search (VNDS) type algorithm in combination with Network Flow Optimization (NFO) models. One main feature of the proposed methodology is that the fitness function used in the GRASP algorithm depends on the result of a LP problem which solves a general covering problem.

The GRASP metaheuristic, introduced by [226] and formally presented by [227], is a multi-start method, with each iteration of the algorithm comprising a constructive phase and a local search phase. The first phase leads to a complete solution, and the second is the improvement phase, which continues until a locally optimal solution is reached. After several iterations of the constructive phase and the local search procedure, the best overall solution is kept as the result. The constructive phase is guided by a greedy function that measures the benefit of including each new element. The benefit of selecting each element changes at each step of the construction. The method is randomized by randomly choosing the next element from a list of candidates. The choice of candidate can be biased by using a family of probability distributions (see [228]). GRASP can be easily hybridized with other approaches and optimization strategies, such as Tabu Search, Simulated Annealing, Variable Neighborhood Search (VNS), and population-based heuristics [229].

The VNS metaheuristic method, introduced by [230], is based on performing systematic changes of neighborhoods during the search space exploration. The application of VNS is quite simple, requiring only the choice of a metric to measure the distance among solutions in the solution space which induces the neighborhood structure. A guide to the application of VNS to various classic problems can be found in [231]. The basic principles of VNS have been extended to provide new versions of the algorithm, which have been successfully applied for solving hard optimization problems. One of the most relevant variants is VNDS which explores neighborhoods in a deterministic way [232].

The choice of neighborhood structure is critical to the performance of a local search algorithm. Basically, the larger the neighborhood, the better the local optimal solutions. However, the larger the neighborhood, the longer it takes to explore. Thus, efficient search procedures are required to get the most out of exploring large neighborhoods. One useful option for exploring very large-scale neighborhoods is to use network flow techniques, as discussed and applied in the context of the travelling salesman and routing problems by [233]. In this and other similar cases (see, for example, [234], [235]), the so-called related graph or improvement graph is a bipartite graph used to represent assignment and matching problems.

The proposed methodology is tested on a real problem by solving the physician scheduling problem in a hospital ER with 42 physicians and a one-year planning horizon. The

mathematical model accounts for all constraints and goals considered in the manual scheduling approach. We show the clear superiority of our hybrid approach over mathematical programming. In fact, in 2018, the solution obtained through the application of the proposed methodology was used in practice for those 42 ER physicians, being deemed by the managers sufficiently superior to replace the manually-created schedule.

The main practical contributions of this and the following chapters are, firstly, to present a mathematical model accounting for all types of constraints and objectives considered in practice by a manager when creating a hospital ER physicians' schedule for a 12-month planning horizon, and secondly, to provide a hybrid algorithm with the capacity to obtain near optimal solutions to large instances of a real physician scheduling problem within minutes. The main methodological contributions of this second part of the thesis are the design of a greedy constructive method with a randomized component dependent upon the exact solution to a general covering problem which is solved by an LP. This latter part of the algorithm provides high quality solutions, in terms both of feasibility and of objective function value (OFV). The proposed VND search method, in combination with NFO, is applied to repair feasibility. Once feasibility is achieved only NFO is used to explore large neighborhoods to improve the OFV. The integration of all these methods provides an algorithm to solve the physician scheduling problem which is efficient and could be adapted to solve other scheduling problems.

Next Section 6.2 summarizes the classification of scheduling problems in literature and the problem addressed in this part of the thesis is identified. Section 6.3 the physician scheduling problem is defined and modelled as an ILP problem.

6.2 Scheduling problem classification

According to Ernst et al. [222] "*Personnel scheduling is the process of constructing work timetables for its staff so that an organization can satisfy the demand for its goods or services.*". This problem is so highly constrained and complex that is very difficult to determine optimal solutions that minimize costs, meet employee preferences, distribute shifts fairly among employees and satisfy all the workplace constraints. These constraints contain rules that regulate the working conditions of personnel and are usually imposed by legislation or company/hospital management: days-off after specific types of shifts, etc.

The scheduling of hospital personnel is particularly challenging because of different staffing needs on different days and shifts due to uncertainty and high fluctuation in the daily requirements for care. Unlike many other organizations, healthcare institutions work around the clock and personnel schedules in practice may sometimes be designed to cope with the demand peaks. This situation leads to irregular shift work, which has an effect on the medical staff well being and job satisfaction impacting upon the working environment and the quality of the delivered service to the patient.

Moreover, the number of qualified personnel required to handle the consistently growing demand for hospital services is rising and workforce related resources represent a large proportion of hospital costs ([215]). Building more efficient personnel schedules for medical staff may reduce these expenses while maintaining quality of care. Meanwhile, modeling preferences and fairness issues may improve work conditions.

The main objectives – some of them already mentioned (e.g., quality of care, workforce expenses, and employee satisfaction) – in physicians scheduling can either by financial or non-financial in nature (or contain both financial and non-financial goals simultaneously). Financial goals are expressed in monetary units while non-financial objectives may focus on individual aspects of physicians and patients and require more sophisticated measures. An example of staff-related targets is fairness in the assignment of shifts, which can be modeled by evenly distributing workloads, including the assignment of unpopular shifts and working hours. Finally, patient-related measures can be direct as for example waiting times or indirect as for example job satisfaction of physicians which have an indirect effect on quality of care.

According to literature [213], [215], [236] medical staff scheduling problems fall into three groups: Staffing, Rostering, and Re-planning problems.

Staffing Problems.

Staffing Problems include strategic decisions concerning the appropriate size and composition of a required workforce, with particular skills, for typically a long-term planning horizon, and the educational program needed by the medical residents. Then, the input data is the demand for services forecasts that the staffing levels need to satisfy by using historical data.

Their main goals are ensuring the coverage of demand in every period in the planning horizon to avoid excessive workload for personnel and long waiting times for patients, to reduce the number of patients who leave the Emergency Department (ED) of a hospital without been treated by medical personnel, etc.

Factors that affect staffing problems include organizational structure and characteristics, personnel recruitment, skill categories of the personnel, working preferences, work agreements for workers. In the current staffing literature there are considerably less percentage of papers that account for individual requests of staff and/or an equal distribution of workload to enhance job motivation than the related percentage of papers in other groups of scheduling problems (rostering and re-planning). Staffing decisions that do not consider fairness may negatively impact on the performance of subsequent rostering and re-planning schedules.

Rostering problems

Rostering problems focus on generating a final schedule that assigns shifts and days off to the individual physicians for a specific planning period. Rostering problems deal with tactical or

operational offline planning decisions. The planning horizon typically spans from weeks to a few months (mid-term problems) in literature.

Tactical planning problems' general goal is to create rosters that are repeated over an extended time horizon to provide predictable schedules for the workforce. Some of the papers such as De Kreuk, Winands, and Vissers's [237] and Gunawan and Lau's [238] study a relatively short cycle time of five weekdays with the intention to repeat these over several months. As these are scheduled during regular working hours, weekends can be neglected in the planning process. Some of the considered objectives are to minimize paid out working hours while meeting demand, deviations from ergonomic, and individual constraints, fairness.

Offline planning problems mainly build detailed schedules for physicians that cover a given time horizon. Their purposes is generally to increase quality of care by reducing over-time hours, granting individual requests, and/or minimizing patient handoffs. Fairness is a key aspect in operational offline rostering problems. For example, Bard, Shu, and Leykum [239] considers fairness in terms of the number of assigned specific shifts.

According to Erhard et al.'s review [215] more than half of rostering publications focus on scheduling staff in the ER, ED, or anesthesia department, as it is the case of the problem we face in this part of the thesis. However, we spans the time horizon to a complete year, which improve work conditions as well as make the problem more difficult to be solved. This difficulty is due to the large number of constraints considered when taking into account specific shift requests and fairness in terms of workload distribution. Furthermore, public holidays and weekends are included, which does not allow the repetition of a relatively short cycle time of several weekdays over the time period.

Re-planning problems.

Re-planning problems deal with short-term decision making as a reaction to unforeseen events, e.g., demand fluctuation and employee absences. According to [215]'s review there is only one paper on this topic in literature ([240]). Operational online planning problems show most potential for future research since current research neglects short-term planning problems such as handling physician absences as well as demand fluctuations.

In reality the three scheduling problems described above often take place on different levels and for completely different time horizons. Interaction between the levels is certainly necessary but in practice it would be unworkable to handle them simultaneously all the time.

Some researchers have decomposed the personnel capacity planning process into four steps [148], [241]–[247]: 1) forecasting demand (based on empirical data), 2) determining staffing requirements over time in order to meet a specific performance target at minimal cost, 3) determining how many workers to assign to each shift type, in order to cover the staffing requirements, and 4) assigning employees to shifts. The three first steps represent the

scheduling problem previously identified as staffing while the forth step represents that identified as rostering. Re-planning problems are not usually considered in the capacity planning but in the online planning problems.

Meanwhile other papers deliberately do not distinguish between the previously defined categories of staffing and rostering, such as the scheduling problems review of Ernst et al. [222] which treats rostering and personnel scheduling as synonymous. They suggest a number of modules associated with the processes of constructing a roster and within these modules different models may be needed for specific applications. These modules start with staff demand modelling and end with the specification of the work to be performed, over a time period, by each individual in the workforce. However, the development of a particular problem may require only some of the modules and, in many practical implementations, several of the modules may be combined into one procedure. Moreover, in an interesting section about decompositions of the problem, one of the described proposals is to present demand modelling as a separate module whose characteristics apply to previously defined staffing problem.

Our scheduling problem, which has already considered to fall within the category of rostering problems, has as input the staff demand based on shifts, which is the possible module 1 according to Ernst et al.'s classification [222]: specification of the number of staff that are required to be on duty during different shifts (redundant to module 3, shift scheduling). This staffing problem is beyond the scope of the problem solved in this part of the thesis, which has to handle the results of management decisions at a higher level.

Meanwhile it does contain other modules such as days off scheduling, line of work construction, and staff assignment. The former involves a determination of how rest days are to be interspersed between work days for different lines of work. The line of work – also called roster lines or work schedules – construction involves the determination of a sequence of duties/shifts spanning the rostering horizon, commonly fortnightly or monthly. These sequences are allocated to individual staff members. The problem usually contains a number of conflicting objectives and constraints. It also considers the rules related to lines of work ensuring the feasibility of individual schedules as well as the pattern of demand satisfying the work requirements at all times in the rostering horizon. In our problem the roster are not cyclic as demand fluctuates with time and where shifts have different lengths and starting times. Our schedule model is what Ernst et al. [222] calls *Line of work constraint* as there are rules governing which work patterns are allowed for an individual such as restrictions on the number of sequential night shifts to be worked, specification of some minimum time off between specific successive shifts, etc. Moreover, an important aspect also faced in our problem is to allow for staff preferences. We generate equitable lines of work that attempt to distribute the workload fairly and evenly within each crew class whilst accommodating each crew class preferences while constructing the lines of work. The latter, staff assignment module, involves the allocation of lines of work to individual staff members belonging to each crew class, which has different preferences and requirements. This assignment in our case is done after generating

all lines of work, as with the bidding systems in which they are then allocated by the department chief.

6.3 Definition and mathematical modelling of the scheduling problem

The solution to the physician scheduling problem lies in determining which physician will work in each shift of each day throughout the planning horizon. Shifts vary in type: there are day and night shifts, workday and holidays shifts, short and long shifts, etc. Even within these categories, there are differences in terms of the task requirement: from the triage area, to the resuscitation room, to consultation for patients with milder symptoms, etc. There is also a variation in the availability and the amount of hours dedicated by the physicians, such as not being able to work all types of shifts. Age or work/life balance issues may prevent certain physicians from working night shifts, for example. Physicians can therefore be grouped by availability and the amount of hours dedicated which means that all members of each group are able to work the same number of hours and types of shift.

The objective of the problem is to obtain the fairest feasible schedule. A fair schedule is one that is evenly distributed among physicians, with all members of a group working the same number of hours, public holidays, weekends, nights (unless exempt), and each type of shift, etc. A balance between groups is also required: the ratio of worked to workable shifts for each physician should be kept proportional across the groups. This workload balancing idea is further developed in Chapter 7.

To offer some idea of the magnitude of this problem, a medium/large size public hospital might have approximately 40 physicians, and approximately 20 different shifts per day. Over a twelve-month planning horizon, this amounts to $365 \times 20 = 7{,}300$ assignments, each with 40 possibilities. The theoretical number of different assignments ($40^{7{,}300}$) is considerably reduced when different types of constraints are included. However, the number of feasible solutions is still huge.

The general formulation of this scheduling problem considers N physicians groupable into M types with n_r physicians of type G_r, $r = 1, \dots, M$, and L types of shifts S_j, $j = 1, \dots, L$, each defined by its duration d_j (in hours), and other characteristics such as night shift, workday shift, the physician's location during the shift, and types of duties required, among others. There are m_j shifts of type S_j in the planning period. Let T be the number of days for the planning horizon.

Each physician type G_r, $r = 1, \dots, M$ can work a maximum of $h_r = \rho_r H$ hours during the planning horizon (where H is the number of working hours of a full time physician and $\rho_r \leq 1$), in a subset of shifts determined by binary indicators γ_{rj}:

$$\gamma_{rj} = \begin{cases} 1 \text{ if a physician of type } G_r \text{can work in shift type } S_j \\ 0 \text{ otherwise} \end{cases} \quad \forall\, r = 1, \dots, M; \forall\, S_j; j = 1, \dots, L \tag{1}$$

Without loss of generality, it is assumed that a subset of shifts $\boldsymbol{S}(t) \subseteq \{S_j, j = 1, \dots, L\}$ needs to be assigned each day and that the demand for each type of shift is one. This assumption reflects the high diversity of shifts in the ER, and places the definition of the problem in a worst case scenario, but the algorithm developed in this research can be straightforwardly adapted for a demand level greater than one. This physician scheduling problem can be mathematically modeled as an ILP problem by using the following decision variables X_{ijt}:

$$X_{ijt} = \begin{cases} 1 \text{ if physician } \mathrm{P}_i \text{ works } \mathrm{S}_j \text{ on day } t \\ 0 \text{ otherwise} \end{cases} \quad \forall \quad i = 1, \dots, N; \quad \forall \quad S_j \in \boldsymbol{S}(t); \quad \forall \quad t = 1, \dots, T \tag{2}$$

Feasible schedules need to cover all shifts, observe the maximum working hours of each physician, and comply with ergonomic constraints (especially those relating to the length of rest period after some types of shifts). Therefore, constraints are classified by type into 1) coverage, 2) ergonomic, and 3) work balance.

- **Coverage constraints.** The demand rules are the most basic compulsory requirements: each physician can be assigned a maximum of one shift per day, and each shift must be assigned to a single physician.

$$\sum_{S_j \in \boldsymbol{S}(t)} X_{ijt} \leq 1 \quad \forall \quad i = 1, \dots, N; \quad \forall \quad t = 1, \dots, T \tag{3}$$

$$\sum_{i=1}^{N} X_{ijt} = 1 \quad \forall \quad S_j \in \boldsymbol{S}(t); \quad \forall \quad t = 1, \dots, T \tag{4}$$

- **Ergonomic constraints.** ER Services are available at all hours of the day and night, every day of the year. Being required to work long shifts at any part of the day without a good distribution of breaks between the shifts turns a poor quality work schedule into a potential health threat for physicians. To mitigate the effects of a chaotic labor calendar, further constraints are added (both to meet legal requirements and accommodate suggestions from physicians) and thus enable physical and mental recovery as well as a normal social and family life. Specifically, these so-called ergonomic constraints, are designed, among other purposes, to avoid

consecutive night shifts, program day(s) off after a long or night shift, plan free weekends, avoid mini-vacation periods between working days, alternate shift lengths, etc.

Ergonomic constraints are classified by purpose into three types: to separate shifts within a specific time, to limit the number of shifts within a time window, and to limit the number of consecutive working days. These constraints can be formulated for each type of shift, for all shifts jointly or for subsets of shifts D_C that share the same characteristics C. For example, D_C ={night shifts worked on public holidays} contains all shifts with characteristics C ={night, public holiday}.

(i) *Minimum days' gap between shifts*. For example, it might be necessary to impose a two-day gap between two worked night shifts (such that there can be only one night shift in a period of 3 days). In this case,

$$\sum_{t=q-2}^{q} \sum_{j \in D_c \equiv \{night\ shifts\}} X_{ijt} \leq 1 \quad \forall \quad q = 3, \dots, T; \quad \forall \quad i = 1, \dots N$$

In general,

$$\sum_{t=q-\delta_c}^{q} \sum_{j \in D_c} X_{ijt} \leq 1 \quad \forall \quad q = \delta_c + 1, \dots, T; \quad \forall \quad i = 1, \dots N \tag{5a}$$

Where D_C is the set of shifts to be interspersed and δ_c is the minimum gap required.

This category of constraints includes a compulsory number of days off after certain types of shifts and is formulated as follows when δ_c days' rest are required after a shift S_{j_c} in a set D_C.

$$\delta_c X_{ij_cq} + \sum_{t=q+1}^{q+\delta_c} \sum_{j} X_{ijt} \leq \delta_c \quad \forall \quad q = 1, \dots, T - \delta_c; \quad \forall \quad i = 1, \dots N \tag{5b}$$

(ii) *Maximum number of shifts worked within a time window*. This type of constraint is used say, to limit the number of public holidays worked within a certain period. Suppose that a physician cannot be assigned more than 5 public holiday shifts over a time window of 30 days. Then

$$\sum_{t=q-29}^{q} \sum_{j \in D_c \equiv \{shifts\ in\ holidays\}} X_{ijt} \leq 5 \quad \forall \quad i = 1, \dots, N: \quad \forall \quad q = 30, \dots, T$$

In general,

$$\sum_{t=q-w_{1c}+1}^{q} \sum_{j \in D_c} X_{ijt} \leq \nu_{1c} \qquad \forall \;\; i = 1, \dots, N; \quad \forall \;\; q = w_{1c}, \dots, T \tag{6}$$

where ν_{1c} is the maximum number of shifts in a set D_c assigned to physicians over a time window of w_{1c} days.

(iii) *Maximum number of consecutive working days*. Physician cannot work more than w_{2c} consecutive days on any type of shift belonging to a set D_c.

$$\sum_{t=q-w_{2c}}^{q} \sum_{j \in D_c} X_{ijt} \leq w_{2c} \qquad \forall \;\; i = 1, \dots, N; \quad \forall \;\; q = w_{2c} + 1, \dots, T \tag{7}$$

Here also, there may be constraints imposing a maximum on the number of days' gap between shifts, a minimum of number of a certain type of shift that can be assigned within a time window, and a minimum on the number of consecutive days on shifts belonging to a set D_c. The formulation of these constraints is similar to that given in (5a) (6) and (7).

- **Workload balancing constraints.** These constraints are designed to guarantee a fair distribution of the different types of shifts among all physicians.

(i) Fair distribution of working hours on shifts belonging to a set D_c among all physicians.

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} d_j X_{ijt} \leq \rho_r H_c^U \qquad \forall \;\; i = 1, \dots, N; \quad \forall \;\; i \; such \; that \; P_i \in G_r \tag{8}$$

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} d_j X_{ijt} \geq \rho_r H_c^L \qquad \forall \;\; i = 1, \dots, N; \quad \forall \;\; i \; such \; that \; P_i \in G_r \tag{9}$$

H_c^U and H_c^L are variables representing the maximum and minimum number of hours worked on shifts with characteristics in C, respectively. These constraints could also be applied to a single type of shifts S_j or to the entire set of shifts.

(ii) Fair distribution among all physicians of shifts in a set D_c

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} X_{ijt} \leq \rho_r J_c^U \qquad \forall \;\; i = 1, \dots, N; \quad \forall \;\; i \; such \; that \; P_i \in G_r \tag{10}$$

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} X_{ijt} \geq \rho_r J_c^L \qquad \forall \;\; i = 1, \dots, N; \quad \forall \;\; i \; such \; that \; P_i \in G_r \tag{11}$$

These constraints are similar to the previous ones, but are now aimed at balancing the number of shifts rather than the number of working hours. The variables J_c^U and J_c^L ,

respectively, limit the maximum and minimum number of shifts worked by all physicians.

(iii) Fair distribution of shifts from a set D_c among physicians in the same group. Constraints for balancing the number of shifts can be assigned to particular types of physicians.

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} X_{ijt} \leq J_{rc}^U \quad \forall \; r = 1, \dots, M; \quad \forall \; i \; such \; that \; P_i \in G_r \tag{12}$$

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} X_{ijt} \geq J_{rc}^L \quad \forall \; r = 1, \dots, M; \quad \forall \; i \; such \; that \; P_i \in G_r \tag{13}$$

The variables J_{rc}^U and J_{rc}^L limit the maximum and minimum number of shifts in set D_c worked by physicians P_i in group G_r, $r = 1, \dots, M$, respectively.

The objective function is defined to reach the fairest distribution of the workload among physicians by minimizing the range of the limiting variables H_c^L and H_c^U, J_c^U and J_c^L, J_{rc}^U and J_{rc}^L. Thus, the objective function is the minimization of the sum of all ranges:

$$\min \sum_{i=1}^{\#D} (H_{c_i}^U - H_{c_i}^L) + \sum_{i=1}^{\#D} (J_{c_i}^U - J_{c_i}^L) + \sum_{i=1}^{\#D} \sum_{r=1}^{M} (J_{rc_i}^U - J_{rc_i}^L), \tag{14}$$

where $\#D$ is the number of sets of shifts D_{c_i} involved in the fairness constraints. Different weights may be used in the objective function to reflect the relative importance of the fairness of the shift distribution and working hours among physicians.

Thus, the ILP model for the physician scheduling problem involves the minimization of the objective function (14) subject to a set of constraints (3)-(13), which is fully presented in Appendix J.

Chapter 7 The hybrid GRASP based algorithm

This chapter explains the hybrid methodology. Section 7.1 provides a general overview of the algorithm. In Section 7.2 a general covering problem is solved by an LP model to obtain the average number of shifts of each type that should be worked by physicians of each type. These averages are used in Subsection 7.3 by a greedy random algorithm to construct a full solution. Finally, Section 7.4 presents two local search procedures to improve the solution obtained by the greedy algorithm.

7.1 General description of the algorithm

The proposed heuristic algorithm comprises three stages: the first solves a global covering and balancing problem formulated as an LP model; the second is a construction phase, in which a full solution is obtained by applying a greedy randomized algorithm (guided by the solution of the first phase); and the third is an improvement stage, in which the solution provided by the previous stage is used as the input to a cyclic optimization alternating between VNDS and NFO which continues until a feasible solution is obtained; this solution is then improved by means of NFO alone. The first stage is executed only once, while the other two stages are iterated several times to define a multi-start procedure, as illustrated in Figure 7.1. This hybrid GRASP-type algorithm will be identified as “Algorithm G+NO”.

The proposed methodology starts by determining the number of each type of shift that each physician should work over the entire planning horizon, in order to guarantee coverage of all shifts and a workload balance among physicians, based on a fair distribution of the different types of shifts (nights, weekends, holidays, etc.). This problem is formulated as a continuous LP problem, which, at a very low computational cost, provides the solution to be used in the next phase.

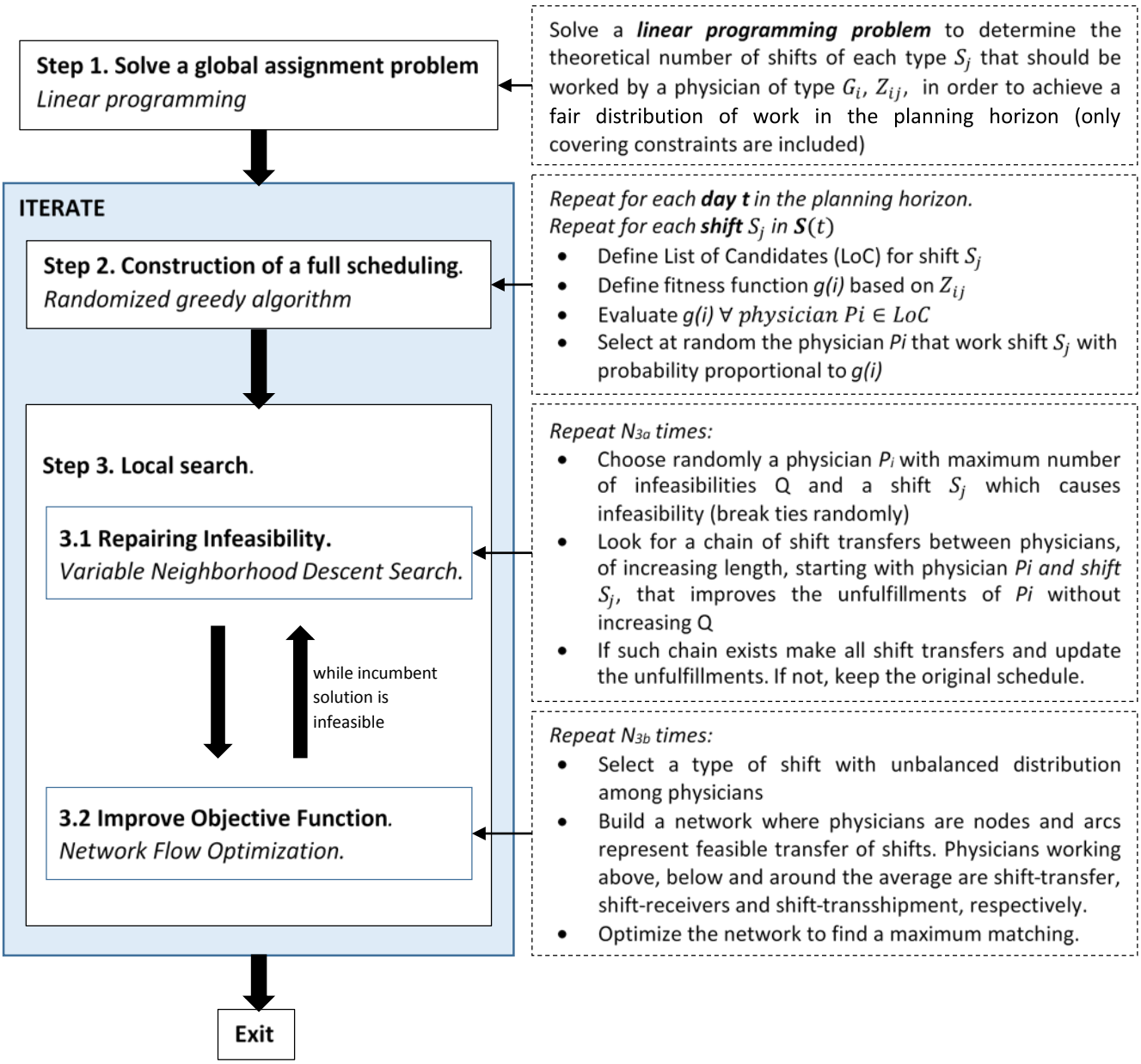


Figure 7.1. The three stages of the proposed heuristic algorithm as applied to physician scheduling.

The construction phase is the implementation of a GRASP algorithm to build a solution by assigning shifts to physicians sequentially. The procedure starts with the first day of the planning horizon, assigning all the shifts for that day and progressing day by day until a full assignment is obtained. The list of candidates for each shift assignment is first defined by the feasibility constraints and then by elitism based on a fitness function. This function takes into account the assignments made so far to all physicians and the theoretical number of shifts of each type that each physician should work (obtained as the solution of the LP formulated in the first phase of the algorithm).

The full scheduling obtained in the previous phase is improved by alternating VNDS to repair violations of the constraints (required if the constructive step provides an infeasible solution) with NFO to balance the distribution of shifts and working hours among the physicians. Once a feasible solution is obtained, improvements to the fair distribution of the workload are sought using NFO only.

In the following subsections, a detailed mathematical and algorithmic description is provided for the three components of the heuristic method.

7.2 A linear programming model to solve the general covering problem

The purpose of this optimization step is to obtain the average number Z_{rj} of shifts of type $S_j, j = 1, \dots, L$, that should be worked by physicians of type G_r, $r = 1, \dots, M$, in order to cover service demand within the regulatory working hours. Variables Z_{rj} can be positive only if $\gamma_{rj} = 1$, that is, $(1 - \gamma_{rj})Z_{rj} = 0$. In addition, this general planning has to distribute the shifts among physicians as evenly and fairly as possible, for which the decision variables Z_{rj} must fulfill the following constraints:

- Demand Covering constraint

$$\sum_{r=1}^{M} n_r Z_{rj} = m_j \qquad \forall \quad S_j;\ j = 1, \dots, L \tag{15}$$

- Working hours constraint

$$\sum_{j=1}^{L} d_j Z_{rj} \leq h_r \qquad \forall \quad r = 1, \dots, M \tag{16}$$

- Equitable distribution of shifts

  Some sets of shifts have to be evenly distributed among those physicians who are able to work them. These include holiday shifts ($D_{hol} = \{shifts\ on\ holidays\}$), night shifts ($D_{nig} = \{shifts\ at\ nights\}$), weekend shifts ($D_{wee} = \{shifts\ on\ weekends\}$), etc.

  Let D_c be the set of shifts to be fairly distributed, and let

$$U_c = \frac{\sum_{S_j \in D_c} m_j}{\sum_r \rho_r n_r \left(1 - \prod_{S_j \in D_c} (1 - \gamma_{rj})\right)}$$

  be the average number of shifts in D_c per full-time physician able to work such shifts. Some shifts belong to one or more sets D_c while others might belong to none. To impose the

equitable distribution of all shifts, two constraints are considered for each set D_c and physician type G_r:

$$\sum_{S_j \in D_c} Z_{rj} - \rho_r U_c \leq F_1 \qquad \forall \ \ r = 1, \dots, M; \ \ \forall \ \ D_c \tag{17}$$

$$\rho_r U_c - \sum_{S_j \in D_c} Z_{rj} \leq F_1 \qquad \forall \ \ r = 1, \dots, M; \ \ \forall \ \ D_c \tag{18}$$

The deviation variable F_1 is minimized in the objective function of the LP problem.

- Even distribution of each type of shift among all physicians. Let

$$W_j = \frac{m_j}{\sum_r \gamma_{rj} \rho_r n_r}$$

be the number of shifts of type S_j that should be worked by each full time physician eligible to do so.

  - Shifts that do not participate in balancing constraints (17) and (18) should also be distributed as fairly as possible. Then,

$$Z_{rj} - \rho_r W_j \leq F_j \rho_r W_j \qquad \forall \ \ r = 1, \dots, M; \quad \forall \ \ S_j \notin \cup_c \{D_c\} \tag{19}$$

$$\rho_r W_j - Z_{rj} \leq F_j \rho_r W_j \qquad \forall \ \ r = 1, \dots, M; \quad \forall \ \ S_j \notin \cup_c \{D_c\} \tag{20}$$

$$F_j \leq F_2^U \qquad \forall \ \ S_j \notin \cup_c \{D_c\} \tag{21}$$

$$F_j \geq F_2^L \qquad \forall \ \ S_j \notin \cup_c \{D_c\} \tag{22}$$

  - Shifts that do participate in balancing constraints (17) and (18) should be distributed as evenly as possible among all physicians.

$$Z_{rj} - \rho_r W_j \leq F_3 \rho_r W_j \qquad \forall \ \ r = 1, \dots, M; \quad \forall \ \ S_j \in \cup_c \{D_c\} \tag{23}$$

$$\rho_r W_j - Z_{rj} \leq F_3 \rho_r W_j \qquad \forall \ \ r = 1, \dots, M; \quad \forall \ \ S_j \in \cup_c \{D_c\} \tag{24}$$

The following objective function (25) minimizes the deviation variables introduced in constraints (17)-(24):

$$\min \beta F_1 + (F_2^U - F_2^L) + F_3 \tag{25}$$

The weighting factor β in the objective function should give much more weight to the first objective than to the others. In fact the optimization can be understood as a lexicographic optimization to find the best proportional shift distributions among all those that are optimal according to the equitable shift distributions in sets D_c. A large enough value for this factor β could be the total number of shifts.

The average number of shifts, Z_{rj}, of each type S_j that should be worked by physicians in group G_r is obtained as the solution of the LP problem with objective function (25) and constraints (15)-(24). The full formulation of the LP problem is included in Appendix K.

For ease of notation, from this subsection forward, the theoretical average number of each type of shift S_j that should be worked by a physician $P_i \in G_r$ will be denoted by Z_{ij}, which is equal to Z_{rj}.

7.3 Construction of a full scheduling solution by a greedy randomized algorithm

This subsection presents a heuristic to generate solutions by a probabilistic greedy construction method. The heuristic follows the constructive step of the GRASP metaheuristic method [248], which builds a solution one element at a time. In the physician scheduling problem, this is done by successively assigning each of the shifts that must be covered each day, starting with a shift from the first day of the planning horizon and ending with a shift from the last day of the planning horizon. Each day's shifts are assigned in random order.

Let T be the number of days in the planning horizon, and $nshifts(t)$ the number of shifts for the t-th day; then the construction phase proceeds in general as shown in Algorithm 7.1:

Initialize $X_{ijt} = 0 \ \forall \ i, \ j, \ t$
for $t = 1$ ***to*** T ***do***
  for $j = 1$ ***to*** $nshifts(t)$ ***do***
    Choose at random a shift S_j not yet assigned
    Define the list of candidates LoC;
    Evaluate each physician in LoC by a greedy function $g: LoC \rightarrow \mathbb{R}$;
    Select a physician $l \in LoC$ by a roulette wheel mechanism
    Add physician l to the set of physicians working on day t, A_t.
  End
End

Algorithm 7.1. Initial solution construction algorithm.

The following subsection gives the details for the definition of the List of Candidates LoC, the definition of the greedy function, and the selection of a physician by a roulette wheel mechanism.

7.3.1 Definition of the List of Candidates

For each assignment of a shift to a physician, a LoC is defined. A physician is included in the LoC for a shift assignment when all the applicable constraints are fulfilled. If the resulting LoC is empty, then all physicians will be included in the LoC. This process is summarized in Algorithm 7.2.

$LoC(j,t) = \{P_i | \text{All constraints for shift } j \text{ are fulfilled}\}$;
If $LoC(j,t) = \emptyset$ ***then***
  $LoC(j,t) = \{P_i, i = 1, \dots, N \,|P_i \notin A_t\}$
End

Algorithm 7.2. List of Candidates for shift j on day t.

7.3.2 Definition of a greedy function $g(i)$

Suppose that a shift of type j has to be assigned on a day t. Let z_{ij}^* be the number of shifts of type j assigned so far to physician P_i and let k be the index of the physician with the maximum value in the following set of ratios:

$$k = \text{arg}max_i \left\{ \frac{z_{ij}^*}{Z_{ij}} \text{ such that } Z_{ij} > 0 \right\} \tag{26}$$

Then, for each physician P_i in the $LoC(j,t)$, the following greedy function g_j(i) is evaluated:

$$g_j(i) = \frac{z_{kj}^*}{Z_{kj}} - \frac{z_{ij}^*}{Z_{ij}} \text{such that } Z_{kj}, \text{Z}_{\text{ij}} > 0 \tag{27}$$

This greedy function measures the difference between the maximum proportion of shifts of type S_j already assigned to a physician (z_{kj}^*/Z_{kj}) and the ratio of shifts assigned to a particular physician. This value is then normalized to the target value for the whole planning horizon, Z_{ij}. Thus, the greater the value of $g_j(i)$ is for physician P_i, the greater his/her need to work this shift S_j in order to meet the reference values Z_{ij}. By definition, this greedy function is a non-negative definite function. However, it could occur that $g_j(i) = 0$ for all physicians in the $LoC(j,t)$.

Enhancement of the greedy function. The greedy function was defined based only on already assigned shifts of type S_j. Nevertheless, some shifts are important for the even distribution of other general shift characteristics among physicians. For example, if the shift that is being assigned is a weekend shift and all physicians have to work the same number of weekends

within the planning horizon; thus, the greedy function must also take into account the consequences of the assignment for the even distribution of weekend shifts. For this purpose, for each set of shifts D_c that has to be evenly distributed among physicians and $S_j \in D_c$, the following greedy function $g_{D_c}(i)$ is defined:

$$g_{D_c}(i) = \frac{\left(\max_l \left\{\frac{z^*_{lD_c}}{Z_{lD_c}}\right\} - \frac{z^*_{lD_c}}{Z_{lD_c}}\right)}{\left(\max_l \left\{\frac{z^*_{lD_c}}{Z_{lD_c}}\right\} - \min_l \left\{\frac{z^*_{lD_c}}{Z_{lD_c}}\right\}\right)} \tag{28}$$

where,

$$Z_{lD_c} = \sum_{S_j \in D_c} Z_{lj} \quad \text{and} \quad z^*_{lD_c} = \sum_{S_j \in D_c} z^*_{lj} \tag{29}$$

A normalized greedy function $g_{Nj}(i)$, which ranges in (0,1), is defined as follows:

$$g_{Nj}(i) = \frac{g_j(i)}{\left(\max_l \left\{\frac{z^*_{lj}}{Z_{lj}}\right\} - \min_k \left\{\frac{z^*_{lj}}{Z_{lj}}\right\}\right)} \tag{30}$$

The new enhanced greedy function $\mathcal{g}_j(i)$ is defined as:

$$\mathcal{g}_j(i) = g_{Nj}(i) + \sum_c g_{D_c}(i) \tag{31}$$

Where the summation is extended to all sets D_c of shifts that need to be balanced and that include the shift S_j.

Then, this greedy function balances the participation of each physician in all shifts and shift characteristics included in the objective function by assigning the shift to the physician who is farthest from meeting all the balancing conditions in which the shift is involved. The balancing assessment takes into account the theoretical values determined by the LP covering problem (Appendix K).

7.3.3 Roulette wheel for the selection of a physician

The probability $p(i)$ of selecting a physician $P_i \in LoC(j,t)$ depends on his/her value in the greedy function:

$$p(i) = \frac{\left(\mathcal{g}_j(i)\right)^\alpha}{\sum_{P_l \in LoC(j,t)} \left(\mathcal{g}_j(l)\right)^\alpha} \tag{32}$$

Observe that if $\alpha = 0$, we will have a random construction; if $\alpha = 1$, the probability will be proportional to the greedy value. The greater the value of α is, the more elitist the selection mechanism.

If all physicians in the $LoC(j,t)$ have $g_j(i) = 0$, then the probability of being chosen is equal among them.

7.4 Improvement of a solution

The feasibility of a solution is improved by decreasing the number of unfulfilled ergonomic constraints by means of a VNDS algorithm, which is followed by a NFO procedure to better fulfill the balancing objectives. These two search mechanisms are applied iteratively (see Figure 7.1) until a stop criterion is met (optimization time or iterations with no improvement). The following subsections offer a detailed description of each of these improvement steps.

7.4.1 Variable neighborhood descent search for repairing infeasibility

The construction phase is driven by the solution of the general covering problem and is particularly oriented towards constructing a feasible solution because the LoC in each shift assignment is first defined by physicians who fulfill all constraints. However, in problems with little slack for finding feasible solutions (too small a surplus with respect to the total demand for working hours and very tough ergonomic requirements), the constructive phase could provide a solution that fails to meet certain constraints. In this case, the first step of the improvement phase is a repair process, whereby a shift contributing to the infeasibility of one physician's schedule is transferred to another physician. These shift transfers successively involve several physicians and are repeated several times. Figure 7.2 represents the logic of these movements: shift S1, which causes the infeasibility of the sequence S1-S2 in physician P_{14}'s schedule (after shift S1, there must be a day off), is transferred to physician P_{23} (causing infeasibility because, two days off are compulsory after shift S7); this requires transferring shift S7 to physician P_9 (again causing an infeasibility), and this, in turn, results in the transfer of shift S3 to physician P_{18}. After these transfers, the initial infeasibility of physician P_{14} is solved without detriment to the total number of non-compliances of the remaining physicians.

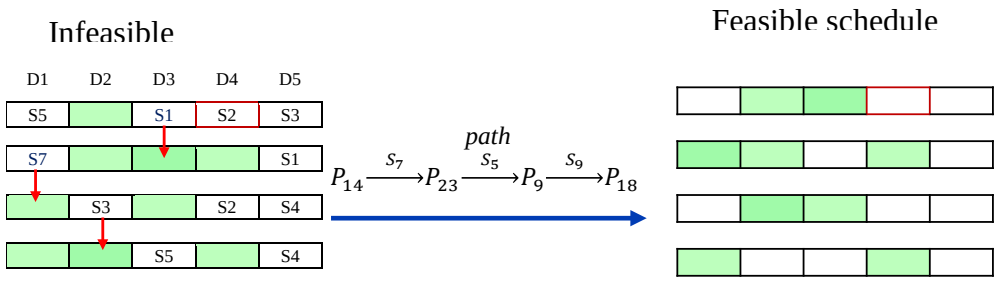


Figure 7.2. Example of shift (S_j: $S1 - S7$) transfer among physicians (P_i: $Physician$ $9, 14, 18, 23$) on different days (D_t: $D1 - D5$). Ergonomic requirements for the different types of shifts: S7 must be followed by two days off; S1, S5 must be followed by one day off; and S2, S3, S4 do not require the next day to be a rest day.

The search for sequences of transfers leading to the improvement of the current schedule falls into the category of a VNDS algorithm with rationale as follows.

Let X_i be the set of shifts assigned to physician P_i in the incumbent solution, that is, $X_i = \{S_j \in \boldsymbol{S}(t) \mid X_{ijt} = 1\}$ and $\rho(X_i, X_i')$ be the distance between solutions for scheduling a physician defined as

$$\rho(X_i, X_i') = |X_i \Delta X_i'| \tag{33}$$

Where $|X_i \Delta X_i'|$ represents the number of shifts that form part of schedule X_i but not of X_i' and those which form part of schedule X_i' but not of X_i. Let us observe that when a physician P_i with schedule X_i transfers a shift to another physician, the resulting schedule for P_i, denoted by X_i', verifies $\rho(X_i, X_i') = 1$.

A full schedule X is the aggregation of all the physicians' schedules: $X = (X_1, X_2, \dots, X_N)$, and then, $\rho(\mathrm{X}, \mathrm{X}') = \sum_{i=1}^{N} \rho\,(X_i, X_i')$ represents the distance between two schedules for all physicians. The transfer of a shift from one schedule X to obtain another schedule X' is denoted by $X' = h(X)$. The schedule solution X' resulting from a sequence of k transfers of shifts in which the transferee in one shift transfer becomes the transferor in the next shift transfer is denoted by $X' = h_p^k(X)$. The index p refers to the path p, which determines the transfers of shifts between physicians. For example, in Figure 7.2, the path is $P_{14} \xrightarrow{S_1\ of\ D3} P_{23} \xrightarrow{S_7\ of\ D1} P_9 \xrightarrow{S_3\ of\ D2} P_{18}$. The length of a path is the number of transfers, in the case of Figure 7.2 the length is 3.

A neighborhood of depth k is defined as

$$\aleph_k(X) = \{X' | \exists\ p \text{ of length } k \text{ such that } X' = h_p^k(X)\}$$

Let us consider a certain type of constraint that is not fulfilled by a solution X and thus requires repair. Let $Q > 0$ denote the maximum number of unfulfilled constraints among all physicians and P_Q the set of physicians that reach this maximum number of non-fulfillments.

$$P_Q = \{\text{physicians with a number } Q \text{ of non} - \text{fulfillments}\}$$

A recursive function enables fairly easy implementation of this VNDS procedure. In each step, each physician with an infeasible schedule tries to transfer a shift (which is problematic because it causes an infeasibility) to another physician, who is able to accept it, even if this results in an additional infeasibility, and then the infeasibility improvement problem is transferred to another physician, and the process is repeated. The steps of this VNDS algorithm are detailed in Algorithm 7.3.

Step 0 Initialize $N_{iter} = 0, k = 0$;

Step 1 $N_{iter} = N_{iter} + 1$;

$k = k + 1$;

if $N_{iter} > max_{iter\_VND}$ **then *End*;**

/* Start new iteration to find a new shift-transfer chain */

Compute set P_Q;

if $P_Q = \emptyset$ **then** Feasible solution, ***End*;**

Choose randomly $P_i \in P_Q$ and shift $S_j \in X_i$ causing a constraint infeasibility;

$S_T \leftarrow S_j$;

$k = 0$;

Go to **Step 2**;

Step 2 $k = k + 1$;

If $k \leq max_{depthSearch}$ **then** /* Begin the exploration of the neighborhood of depth k*/

if $\exists\, Physician\ P_{i^*} \mid X'_{i^*} = X_{i^*} \cup \{S_T\}$ *does not increase the infeasibilities of* P_{i^*} **then**

Make definitive all temporal transfers and go to **Step 1**;

else if $\exists\, Physician\ P_{i^*} \mid X'_{i^*} = X_{i^*} \cup \{S_T\}$*does not increase the value of* Q **then**

Transfer temporarily shift S_T to P_{i^*};

Select shift $S_j \in X_{i^*}$ $(S_j \neq S_T)$ that causes constraint infeasibility to P_{i^*};

$i \leftarrow i^*, S_T \leftarrow S_j$;/* Solving the infeasibility problem is transferred from P_i to P_{i^*}*/

Go to **Step 2**;

else

Undo the temporally transfers, keeping the initial schedule and go to **Step 1;**

end

end

Algorithm 7.3. VNDS Procedure for repairing solutions. It is based on transferring a shift contributing to the infeasibility of one physician's schedule to another physician. These transfers of shifts successively involveseveral doctors and are repeated several times.

7.4.2 A network flow optimization problem for balancing the distribution of shifts and working hours

The goal of this optimization procedure is to transfer shifts from physicians with surplus, working significantly more than average hours to physicians with slack in that type of shifts, and working significantly fewer than average hours. The term "significantly" is used in relation to a zone of indifference surrounding the average number of hours worked, which is defined in order to stabilize the procedure as it progresses. A physician P_i is considered to have an acceptable total of working hours, $H_i(X) = \sum_{t=1}^{T} \sum_{j=1}^{L} d_j X_{ijt}$, in a schedule X when as long as it belongs to this interval of indifference. To formalize this idea, for each iteration l of this optimization procedure $\left(1 \leq l \leq \max_{iter\_NFO}\right)$, the lower and upper boundaries of the indifference interval, LH and UH respectively, around the average number of working hours are defined as follows:

$$LH = \rho_r \bar{H} \left(1 - \left(\frac{l}{\max_{iter\_NFO}}\right) \varepsilon\right) \tag{34}$$

$$UH = \rho_r \bar{H} \left(1 + \left(\frac{l}{\max_{iter\_NFO}}\right) \varepsilon\right) \tag{35}$$

Where ε is the factor defining the final window of indifference. For example, $\varepsilon = 0.0015$ and an average $\bar{H} = 1{,}750$ and a full-time physician ($\rho_r = 1$); the indifference window is $UH - LH \approx 5\ hours$. The average $\bar{H}$ for a full-time physician can be estimated as $\bar{H} = \frac{\sum_j d_j m_j}{\sum_r \rho_r n_r}$

Given a schedule solution X, these two limits classify the physicians into three groups:

$P_{TS}(X) = \{P_i \mid H_i(X) > UH\}$

$P_{RS}(X) = \{P_i \mid H_i(X) < LH\}$

$P_{IN}(X) = \{P_i \mid LH \leq H_i(X) \leq UH\}$

The physicians in set $P_{TS}(X)$ can transfer shifts, and those in set $P_{RS}(X)$ can receive shifts. Physicians in the balanced set $P_{IN}(X)$ can play an intermediate role by both receiving and transferring shifts. This condition for transferring a shift is called the working hours' condition (WHC).

A physician of type G_r can transfer a shift of a certain type S_j when the number of assignments of this type exceeds the theoretical number Z_{ij} determined in the pre-processing optimization phase; and, conversely, a physician can receive a shift of a certain type when the number of assignments of this type is below this theoretical figure. In terms of the notation introduced in Section 7.2, a physician P_i is allowed to transfer a shift S_j when $z^*_{iD_c} > Z_{iD_c}$, and a physician P_i is allowed to receive a shift S_j when $z^*_{iD_c} < Z_{iD_c}$ for all sets D_c with relevance in the objective function and in which shift S_j participates. This shift transfer condition is named the balancing shift condition (BSC).

Building the network structure. The nodes represent physicians, and each arc (i, k) represents a possible transfer of a shift S_j from physician P_i to physician P_k. The physician P_i belongs to set $P_{TS}(X)$, and P_k belongs to set $P_{RS}(X) \cup P_{IN}(X)$, or P_i belongs to set $P_{IN}(X)$, and P_k belongs to set $P_{RS}(X)$. To plot an arc on the graph, both physicians, transferor and transferee, must meet the conditions WHC and BSC defined earlier and the transferee must be feasibly able to work this shift. When there exists more than one arc verifying the conditions between a pair of physicians, one of them is chosen at random (since it is the case that more than one shift could feasibly be transferred from physician P_i to physician P_k). Therefore, the network structure is built randomly and successive iterations of this procedure provide different networks.

Assigning demands, capacities, and costs to the network. Nodes representing a physician in $P_{TS}(X)$ have a demand of -1, nodes representing a physician in $P_{RS}(X)$ have a demand of $+1$, and nodes representing a physician in $P_{IN}(X)$ have a demand of 0 (trans-shipment nodes).

The network is expanded by unfolding each node in the set $P_{IN}(X)$, into two nodes that are connected by an arc.

All arcs in the network have a maximum capacity of 1 and a minimum capacity of 0.

Costs:

- the arcs between a physician in $P_{TS}(X)$ and a physician in $P_{RS}(X)$ have a cost of -2,
- the arcs between a physician in $P_{TS}(X)$ and a physician in $P_{IN}(X)$, or between a physician in $P_{IN}(X)$ and a physician in $P_{RS}(X)$ have a cost of -1,
- the arcs between nodes representing the same physician in $P_{IN}(X)$ have a cost of 0.

Figure 7.3 shows a simple example of a flow network with 3 physicians in set $P_{TS}(X)$, 3 physicians in set $P_{IN}(X)$, and 2 physicians in set $P_{RS}(X)$.

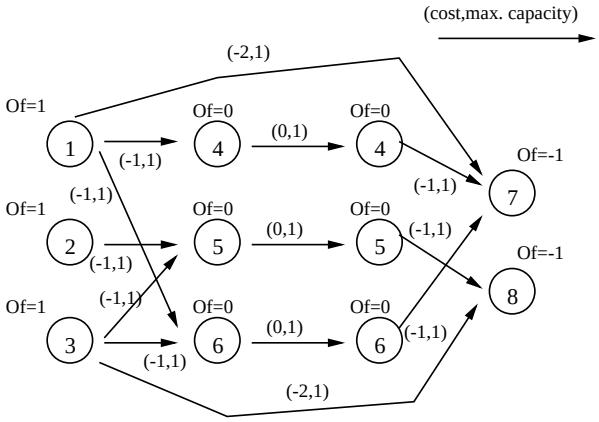


Figure 7.3. Example of work-flow network. Physicians 1, 2 and 3 can transfer one shift; physicians 4, 5 and 6 can receive and transfer one shift, and physicians 7 and 8 can receive one shift.

Solving the network flow problem. The problem is solved by using an algorithm to find the minimum-cost feasible flow. The resulting networks are small in size and can be solved quickly by efficient algorithms such as Network Simplex, Out of Kilter, Cycle Canceling, or Successive Shortest Path (see [249]). Our implementation uses a successive shortest path algorithm, as described in [250]. After network optimization, each physician can transfer and receive, at most, one shift. For this reason, this optimization step is repeated $\max_{iter\_NFO}$ times. In each iteration, the limits that define the partition of physicians into sets $P_{TS}(X)$, $P_{RS}(X)$, and $P_{IN}(X)$ are modified, starting with small values, which are gradually increased. Any fluctuation of the zone of indifference between two values contributes to the variability of the created networks and the stabilization of the shift transfers as the algorithm progresses.

Consecutive iterations of this procedure lead to different networks, which gradually improve the balancing of shifts and working hours. When this NFO phase is iterated with the VNDS algorithm because the solution is still infeasible, the NFO helps the VNDS algorithm by providing new starting solutions from which to search for good shift transfer chains (as in a shaking procedure) and also helping to redress any imbalance in the shift distribution that may be introduced due to the application of VNDS.

Chapter 8 Computational analysis

This section reports the results of the empirical assessment of the algorithm presented in previous Chapter 7. Its practical effectiveness is tested in Section 8.1 by solving the problem of scheduling all the ER shifts for the year 2018 among 42 physicians in the Hospital Complex of Navarre (HCN) in Spain. In addition, in Section 8.2, a set of synthetic scheduling problems of varying degrees of difficulty is used to assess the performance of the algorithm under different conditions. The results are compared with those obtained by CPLEX. Section 8.3 investigates the influence of the different phases of the algorithm on the solutions to the physician scheduling problem as well as the value of its parameters to obtain good solutions while Section 8.4 describes the implementation of the algorithm. Finally the chapter ends with the conclusions about this whole part of the thesis (Chapter 6, Chapter 7, and Chapter 8).

8.1 The physician scheduling problem at the Hospital Complex of Navarre (HCN)

The ED of the HCN, which is more detailed described in Chapter 2, is staffed 24 hours per day by 42 board-certified emergency physicians. Currently, each year's shift schedule is planned manually by one of the physicians, who dedicates three weeks' work to this task. Although, this person is an experienced physician and has been in charge of schedule planning for many years, the task becomes more complicated every year, because new labor laws create new constraints and new categories of workers with different working conditions. This physician creates the schedule without any technological/computational support, using only large spread sheets, similar to the one shown in Figure 8.1, where there is a row for each physician and a column for each day. Starting with simple rotational rules, the scheduler uses his/her own heuristics to consecutively balance holiday shifts, weekend shifts, nights, and, finally, regular shifts, while also trying, to satisfy a large set of constraints (ergonomic, workload, etc. as described in Section 6.3). The resulting schedule violates many conditions as well as provoking numerous complaints from other physicians, who consider the schedule unbalanced and conditioned by subjective preferences.

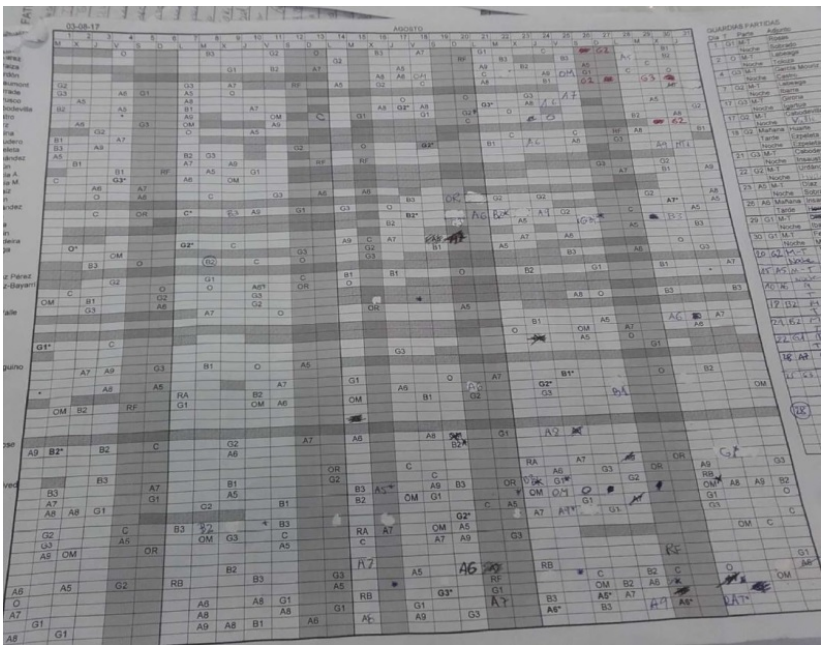


Figure 8.1. The hospital's current scheduling method.

Staff characteristics. The staff comprises 42 physicians who can be grouped into two types: 1) a first group G_1 of 3 physicians who are exempt from night shifts (denoted by O, A5, G1, G2 and G3 in Table 8.1) for reasons of age or various other reasons such as work-family reconciliation and 2) a second group G_2 of 39 physicians who can work any shift.

Shift characteristics. Shifts differ in length and task characteristics. In the ER of the HCN, physicians can be assigned to different areas, such as the resuscitation room, the triage zone, the observation zone, or the severe patient circuit. Each of these locations involves different tasks and responsibilities. In addition, different numbers and types of shifts are scheduled for different types of days. Table 8.1 includes relevant information about shift length, the type of shifts worked per type of day, and the number of days off after each shift. A balanced distribution of all types of shifts among the physicians must be achieved.

Table 8.1. Shift coverage requirements by type of day. The shift labels (S1-S19) are those used by the ER of HCN (row 2: local description).

Shifts	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19
Local Description	G1	G2	G3	A5	O	C	B1	B2	B3	A6	A7	A8	A9	OM	R	OR	RF	RA	RB
Length in hours	19	19	19	19	20	14	14	14	14	14	14	8	8	8	3	14	14	14	14
Workdays	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X				
Mondays*	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			X	X
Holidays	X	X	X	X	X	X				X	X					X	X		
Days off after shift	2	2	2	2	2	1	1	1	1	1	1	0	0	0	0	1	1	1	1

*Mondays or any other day following a holiday.

Constraints. There are some compulsory requirements for individual schedules: two days off have to be scheduled after a long shift (19/20 hours) and one day off after a 14-hour-shift; schedules must not allow more than two consecutive weekend shifts; or more than 5 holiday shifts in a month (these include Saturdays and Sundays); and must allow a four-day gap between night shifts. In addition, all physicians' schedules must fulfill certain balanced distribution criteria based on the number of shifts of each type worked yearly (13 balance conditions, B1 to B13, defined in Table 8.2), and all these shifts have to be evenly distributed over the year.

Table 8.2. The 13 balancing objectives.

Balancing objective name	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13
Type of shifts to be balanced	C	B1+B2+B3	A6+A7	A8+A9	OM	(A6+A7)--(OR+RF)	O	A5	G3	G1+G2	D_C = = {holidays}	D_C = = {weekends}	AWH

Results. The problem was first formulated as a Mixed Integer Linear Programming model (see Appendix J) with over 200,000 variables and 70,000 constraints. CPLEX 12.6.2 solved this problem on an Intel (R) Xeon (R) CPU E5-1630 v4 3.70GHz and 64.0 GB RAM, and after an entire week of execution time, the best-found integer solution provided an objective function value of 43 (see Table 8.3), which was obtained after 168 computation hours and remained unchanged for 54 hours, until the end of the experiment (see Figure 8.5). However, CPLEX was not able to prove optimality of that best-found solution within the computational time limit. In fact, it is not optimal, because the G+NO algorithm obtained a solution with a OFV of 15 within seconds. Figure 8.5 shows the best-found solutions obtained by both CPLEX and the G+NO algorithm over time (note that the time axis is expressed in seconds for the G+NO algorithm and in hours for CPLEX).

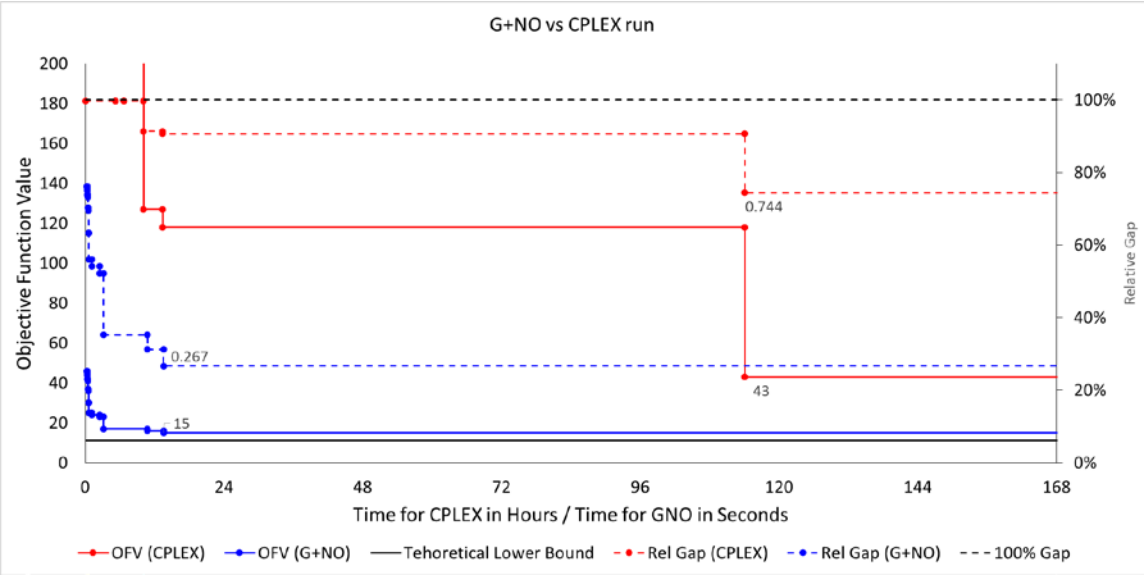


Figure 8.2. CPLEX and G+NO algorithm performance: best found solutions obtained by both over time.

To apply the G+NO algorithm, the initial LP problem was first formulated in order to obtain the optimum theoretical values of Z_{ij} for each type of shift and physician group $Z_{ij}, i = 1, 2; j = 1, \dots, 19$, which was solved within seconds. Table 8.3 shows, in row 3, the theoretical optimum value for each shift-balancing goal $B_k, k = 1, \dots, 13$ for the two groups of physicians. These values guided the constructive phase and the objective function improvement. The maximum and minimum numbers of shifts worked by a physician in either group according to the solution obtained by CPLEX, in one hour and in one week, are given in rows 4-5, and 6-7, respectively; and in rows 8-9 for a G+NO solution obtained after five minutes' computation time. The column for B13 shows the hours worked annually, and it is here that the G+NO clearly outperforms CPLEX, thus demonstrating the efficacy of the Network improvement phase. The best bound obtained by CPLEX in one week is 4.547. A straightforward analysis of the objective function can provide better bounds; superior to those provided by CPLEX (see Table 8.3).

The notion behind this target bound is the following: when the number of shifts participating in a balancing goal is not a multiple of the number of possible shift candidates, it is impossible for them all to be assigned the same number of shifts of this type, and the balanced solution will, therefore, necessarily fall within a range of at least one. However, when the number of shifts is a multiple of the number of candidates then an even distribution among all physicians is possible. This simple analysis provides a minimum bound for the objective function. In the case study, this bound is 11 and G+NO and CPLEX solutions provide a relative gap (36) of 0.27 and 0.74, respectively.

$$GAP = \frac{\{OFV\} - \{Theoretical\ bound\}}{\{OFV\}} \tag{36}$$

The solution obtained with the heuristic obtains the bound for each balancing criterion except for B6, which could theoretically obtain a value of 0 but in fact obtains a range of 1; criterion B11, which could theoretically obtain a value of 0 and actually obtains a range of 2; and criterion B13, which could theoretically obtain a value of 1 and actually obtains a range of 2. These differences increase the global bound of 11 by 4 units to an OFV of 15. In conclusion, the solution may be non-optimal, but, from a practical point of view it is, nevertheless, a very high quality solution compared with those obtained manually by the physician, who accepted solutions within a range of 2 or 3 for goals B1-B12 and a range of 20 for goal B13.

Table 8.3. Case study results: heuristic algorithm and CPLEX results for balancing the different shift sets (B1-B13) included in the objective function OFV. Max. and Min. refer to the maximum and minimum number of balancing goals involving physicians in the corresponding group. The relative gap (last column is calculated relative to the theoretical minimum bound).

		Obj G1						Obj G2										Obj G1&G2				
Objectives		$B1_1$	$B2_1$	$B3_1$	$B4_1$	$B5_1$	$B6_1$	$B1_2$	$B2_2$	$B3_2$	$B4_2$	$B5_2$	$B6_2$	$B7$	$B8$	$B9$	$B10$	$B11$	$B12$	$B13$	OFV	$Rel. Gap$
Theoretical bound		**17.38**	**35**	**32.59**	**23.33**	**11.67**	**0**	**8**	**16.15**	**16.21**	**10.77**	**5.38**	**0**	**9.36**	**9.36**	**9.36**	**18.72**	**25**	**3.57**	**1750.95**	**11**	**0**
CPLEX (1hour)	Max.	174	133	41	0	0	41	51	46	44	62	138	19	52	43	56	70	36	8	2515	3451	1
	Min.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78		
CPLEX (1week)	Max.	25	31	36	23	7	1	8	20	17	12	7	3	10	11	10	21	4	26	1756	43	0.74
	Min.	25	31	36	23	7	0	7	15	14	10	4	0	8	8	8	17	3	24	1745		
G+NO (5min)	Max.	17	35	33	24	12	0	9	17	17	11	6	1	10	10	10	19	26	4	1752	15	0.27
	Min.	17	35	33	24	12	0	8	16	16	10	5	0	9	9	9	18	24	3	1750		

8.2 Additional computational experiments

In this section, the performance and efficacy of the proposed algorithm are evaluated by creating new instances in order to obtain problems of different sizes and degrees of difficulty, while still maintaining the characteristics of a real problem. From the real case detailed in Section 8.1, two more different-sized problems with 20 and 30 physicians, respectively, were created by rescaling all the physician and shift types included in the real case. These three instances (the real case with 42 physicians and the two new rescaled instances with 20 and 30 physicians, respectively) are considered normal-difficulty instances, and highlighted in bold in Table 8.4.

Eight more instances, all for different sized problems, were designed. Four of them are intended to increase the solving difficulty by increasing the number of shifts to be assigned in total and therefore per physician, thus making the ergonomic constraints more difficult to satisfy. The other four scenarios are designed to facilitate the process by decreasing the number of shifts. Specifically, the new problems are obtained as follows:

- The less difficult instances. The number of shifts assigned per day is reduced by one on some days to obtain the four new problems: workday morning shift, holiday morning shift, workday night shift, and holiday night shift, respectively. Thus, the ratio of average working hours (AWH) with respect to the initial scenario is less than one.
- The more difficult instances. The number of shifts assigned per day is increased by one on some days to obtain the four new problems: workday morning shift, holiday morning shift, workday night shift, and holiday night shift, respectively. Thus, the ratio of average working hours (AWH) with respect to the initial scenario is greater than one.

The increases and reductions in the number of shifts can also change the number of holidays and number of nights worked by a physician, thereby affecting the difficulty of solving the problem. Table 8.4 compares the results of all instances provided in 5 minutes by the heuristic algorithm and in one hour by CPLEX on the same computer. The table includes the ratios of physicians, holidays, worked nights, and annual hours worked per physician in each solved

instance with respect to the reference problem. The results provide the objective function value (OFV), the range of annual hours worked (explicitly included because of the difficulty involved in balancing it) and the gap with respect to the theoretical bound. The best bound obtained by CPLEX and the gap with respect to it is also included. The heuristic G+NO algorithm is run 30 times for 5 minutes each. The heuristic algorithm is a multi-start algorithm, set to generate 10 solutions and improve them for a total of 30 seconds each (easiest problems with $AWH < 1$), or 5 solutions with an improvement time of 1 minute (harder problems with $AWH \geq 1$). The algorithm returns the best of these 5 or 10 solutions. Table 8.4 presents the results for the best of the 30 runs, the average solution and the median solution. The heuristic algorithm outperforms CPLEX in all instances: the mean and median of the 30 runs of the heuristic algorithm are much lower than the OFV obtained by CPLEX. In all instances, moreover, the 30 runs of the G+NO algorithm provide a better solution than the CPLEX.

Table 8.4. Comparison of the solution obtained by the heuristic algorithm in five minutes with the one provided by CPLEX in one hour.

No. of Physicians (Phys)	Instances Description (standard = 1)			Theoretical Bound	**CPLEX** performance			**CPLEX** solution		**G+NO** performance (BEST)			**G+NO** performance (AVERAGE)			**G+NO** performance (MEDIAN)			Improvement over CPLEX
	Holidays per Phys ratio	Nights per Phys ratio	AWH ratio		OFV	AWH	Rel. Gap	Best Bound	Rel. Gap	OFV	AWH	Rel. Gap	OFV	AWH	Rel. Gap	OFV	AWH	Rel. Gap	
20	0.80	0.85	0.93	7	20	9	0.65	1.87	0.91	8	2	0.13	10.50	3.40	0.33	11	3	0.36	0.60
20	1.00	0.79	0.91	8	16	7	0.50	4.52	0.72	9	2	0.11	10.27	2.13	0.22	10	2	0.20	0.44
20	0.80	1.00	0.95	7	15	7	0.53	1.22	0.92	8	1	0.13	9.03	2.10	0.23	9	2	0.22	0.47
20	1.00	1.00	0.98	9	23	6	0.61	3.88	0.83	10	1	0.10	10.27	1.07	0.12	10	1	0.10	0.57
20	1.00	1.00	1.00	**9**	**24**	**6**	**0.63**	**3.03**	**0.87**	**11**	**2**	**0.18**	**12.20**	**3.40**	**0.26**	**12**	**3**	**0.25**	0.54
20	1.00	1.00	1.07	8	41	6	0.80	2.88	0.93	11	3	0.27	14	5.17	0.43	13.5	5	0.41	0.73
20	1.20	1.00	1.05	9	32	7	0.72	1.34	0.96	12	2	0.25	14.17	4.73	0.36	14	5	0.36	0.63
20	1.00	1.21	1.09	9	30	7	0.70	1.26	0.96	14	4	0.36	16.67	5.87	0.46	16	5.5	0.44	0.53
20	1.20	1.15	1.07	9	23	14	0.61	0.75	0.97	11	2	0.18	17.63	6.20	0.49	16.5	5	0.45	0.52
30	0.86	0.91	0.95	9	67	49	0.87	0.52	0.99	12	2	0.25	14.47	3.13	0.38	14	3	0.36	0.82
30	1.00	0.87	0.94	12	63	9	0.81	3.18	0.95	14	2	0.14	14.97	3	0.20	15	3	0.20	0.78
30	0.86	1.00	0.97	9	22	11	0.59	0.32	0.99	12	1	0.25	14.87	2.23	0.39	15	2	0.40	0.45
30	1.00	1.00	0.97	11	26	13	0.58	1.34	0.95	14	2	0.21	15.53	3.23	0.29	15.5	3	0.29	0.46
30	1.00	1.00	1.00	**11**	**25**	**12**	**0.56**	**0.63**	**0.97**	**13**	**1**	**0.15**	**14.00**	**1.83**	**0.21**	**14**	**2**	**0.21**	0.48
30	1.00	1.00	1.03	11	3151	2402	1.00	1.59	1.00	14	1	0.21	15.43	3.03	0.29	15	3	0.27	1.00
30	1.14	1.00	1.03	9	2355	1932	1.00	1.21	1.00	13	1	0.31	15.47	2.2	0.42	15.5	2	0.42	0.99
30	1.00	1.13	1.06	11	3077	2379	1.00	1.28	1.00	14	3	0.21	15.30	3.1	0.28	15	3	0.27	1.00
30	1.14	1.09	1.04	10	3038	2359	1.00	1.06	1.00	14	1	0.29	16.13	1.9	0.38	16	2	0.38	1.00
42	0.90	0.93	0.97	12	3387	2379	1.00	0.00	1.00	14	1	0.14	15.53	2.83	0.23	16	3	0.25	1.00
42	1.00	0.91	0.96	11	3352	2427	1.00	0.00	1.00	15	3	0.27	16.27	3.30	0.32	16	3	0.31	1.00
42	0.90	1.00	0.98	12	3309	2379	1.00	0.00	1.00	14	1	0.14	15.67	1.73	0.23	16	2	0.25	1.00
42	1.00	1.00	0.97	11	3129	2379	1.00	0.00	1.00	14	3	0.21	15.67	3.23	0.30	16	3	0.31	1.00
42	1.00	1.00	1.00	**11**	**3451**	**2437**	**1.00**	**0.00**	**1.00**	**15**	**1**	**0.27**	**16.43**	**1.87**	**0.33**	**16**	**2**	**0.31**	1.00
42	1.00	1.00	1.02	11	3516	2437	1.00	0.00	1.00	14	2	0.21	16.77	3.00	0.34	17	3	0.35	1.00
42	1.10	1.00	1.01	13	3453	2379	1.00	0.00	1.00	16	3	0.19	17.47	3.50	0.26	18	3	0.28	1.00
42	1.00	1.09	1.04	10	3292	2379	1.00	0.00	1.00	14	2	0.29	15.47	3.53	0.35	16	4	0.38	1.00
42	1.10	1.07	1.03	12	3414	2413	1.00	0.00	1.00	15	2	0.20	16.60	2.93	0.28	17	3	0.29	1.00

Observe that, in problems with 20 physicians and fewer/weaker constraints (first four scenarios), the best G+NO solution is only one unit's distance from the theoretical bound, and in all scenarios this distance is less than or equal to 4, except in one where it is 5. As already mentioned, these results are very good from a practical point of view, since they considerably improve the manually designed schedules which were not feasible and had wider-ranging balancing criteria.

The quality of each solution is assessed by the gap (36) with respect to the bound calculated from the theoretical analysis, which is included in Table 8.4. Column of improvement over CPLEX is calculated similarly (37):

$$\text{Improvement} = \frac{\{CPLEX\ OFV\} - \{Best\ G + NO\ OFV\}}{\{CPLEX\ OFV\}} \tag{37}$$

8.3 Parameter tuning

In this section we investigate the influence of the different phases of the algorithm and the value of its parameters for obtaining good solutions to the physician scheduling problem. We deal with the capacity of the algorithm first to achieve feasible solutions and then to improve the value of the objective function.

Fine-tuning of parameters to obtain feasible solutions. The constructive phase of the algorithm includes feasibility as the first condition for defining the LoC from which a physician will be selected at random to be assigned a shift. Thus, in problems with no heavy constraints, the constructive phase is expected to provide a feasible solution. However, this does not occur in problems heavily constrained by strict ergonomic requirements and heavy workloads. To illustrate this, we conducted an experiment using the 27 problems solved in the previous section, obtaining, for each one, 100 different solutions using only the constructive phase of the algorithm. Table 8.5 contains the number of feasible solutions. Clearly, when one extra holiday and night shift are added, and there are fewer physicians to share the extra work, the problem becomes harder to solve. However, when feasibility is not achieved, the number of infeasibilities is low, usually one or two (out of the several tens of thousands of constraints). In the case of the 20-physician problem, with one night shift added on every holiday, none of the 100 solutions provided by the constructive phase is feasible. In this worst-case scenario, the number of infeasibilities could reach around 10-15 (Figure 8.3 shows the distribution of the number of unfulfilled constraints in the one hundred solutions of the two worst instances: when an extra night shift or an extra day shift are added on holidays for an ED with 20 physicians). In instances with no heavy constraints, the constructive phase obtains a feasible solution within 100 runs.

Table 8.5. Percentage of feasible solutions reached in the constructive phase of the G+NO algorithm.

Instances									
Shifts added		0	+1	+1	+1	+1	-1	-1	-1	-1
Time-slot			Day	Night	Day	Night	Day	Night	Day	Night
Type of day			Holiday	Holiday	Work day	Work day	Holiday	Holiday	Work day	Work day
Nº. physicians	20	100	28	0	100	84	100	100	100	100
	30	99	89	19	99	81	100	100	100	100
	40	100	99	92	100	99	100	100	100	100

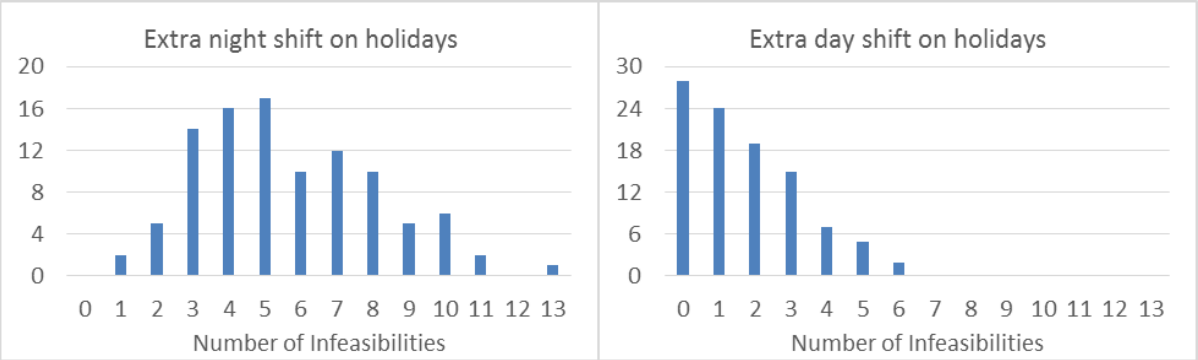


Figure 8.3. Distribution of the number of infeasibilities for the two hardest scenarios (they both have 33379 constraints).

To analyze the performance of the feasibility improvement phase, we use the most difficult problem that of scheduling shifts for 20 physicians, for which no feasible solution was obtained initially. Specifically, we study the influence of two parameters: the number of iterations max_{iter_VND} of the VNDS algorithm; and the number of iterations max_{iter_NFO} of the NFO step. The recursion depth parameter is set as 10, which is large enough to permit a wide search and small enough to avoid excessive memory consumption (higher values can lead to memory allocation problems).

For each combination of the values 1, 5, 10, 20 and 50 for max_{iter_VND} and 1, 5, 25, 50, 100, and 200 for max_{iter_NFO}, 50 solutions are obtained by running the algorithm G+NO for 30 seconds. Thus, 1,500 different solutions are obtained for the same problem. Table 8.6 shows the percentage of feasible solutions obtained with each combination of parameters. A two-way ANOVA reveals the influence of the value max_{iter_NFO} in the results (p-value<0,001) but not the influence of max_{iter_VND} (p_value=0,915). The results of a post-hoc analysis of a one-way ANOVA, using only max_{iter_NFO}, and the graph of means (Figure 8.4) reveals that results for 1 and 5 are much worse and that significantly better results are obtained for values of 25 and 50 (after which they deteriorate slowly as the number of iterations increases). An explanation for these results is the following: given a schedule, the VNDS tries to sequentially find shift-

transfer chains to repair infeasibilities; but, in heavily constrained problems, it is possible that no (or only very few) such chains exist in the current solution. Therefore, it is necessary to shake the current solution to obtain a new one and then resume the search for the required feasibility-repairing chains. These new schedules are provided by applying the NFO step. The results show that too few iterations (1-5) do not create significantly different solutions, whereas too large values of max_{iter_NFO} consume extra computational time. From this analysis, values of around 25-50 could be considered appropriate.

Table 8.6. % of feasible solutions reached by algorithm 4 for each configuration.

		max_{iter_VND}				
		1	**5**	**10**	**20**	**50**
max_{iter_NFO}	**1**	4	6	12	10	4
	5	26	24	28	22	14
	25	78	82	82	90	82
	50	80	78	88	72	78
	100	72	74	74	74	84
	200	76	66	62	78	82

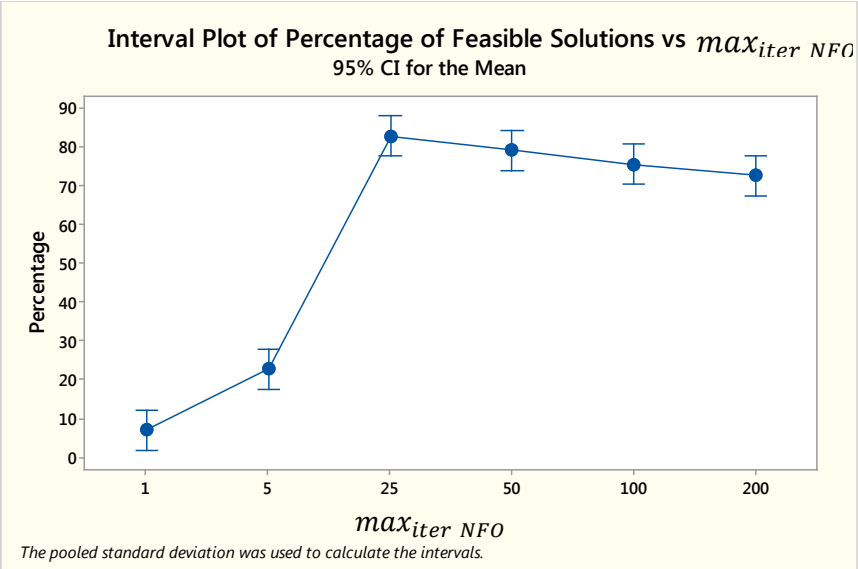


Figure 8.4. Graph of the mean % of feasible solutions reached by algorithm for each max_{iter_NFO} parameter value.

Fine-tuning to obtain good objective function values. Table 8.7 shows the average, median, and minimum of the feasible solutions obtained after running the algorithm 50 times for one minute for each combination of max_{iter_NFO} and and max_{iter_VND} parameters. We consider in the study the values for max_{iter_NFO} identified as acceptable in Figure 8.4 for obtaining feasible solutions (25, 50, 100) and the values 1, 5, 10, 20, 50 for max_{iter_VND}. The results show no statistically significant differences. However, in order to fix parameter values, we choose 25 for max_{iter_NFO} and 20 for max_{iter_VND}, because they provide the lowest mean, median, and minimum values.

Table 8.7. Mean, median and minimum values of the 50 iterations of G+NO algorithm.

MEAN		max_{iter_VND}				
		1	**5**	**10**	**20**	**50**
max_{iter_NFO}	**25**	22.76	21.46	20.93	19.84	20.10
	50	24.12	20.19	25.33	23.82	22.52
	100	20.67	23.82	22.55	20.63	22.28

MEDIAN		max_{iter_VND}				
		1	**5**	**10**	**20**	**50**
max_{iter_NFO}	**25**	19.0	19.0	18.0	17.5	18.0
	50	22.0	19.0	23.0	22.0	21.0
	100	18.5	21.5	21.0	18.0	19.5

MINIMUM		max_{iter_VND}				
		1	**5**	**10**	**20**	**50**
max_{iter_NFO}	**25**	13	13	13	12	13
	50	13	13	13	12	12
	100	12	13	13	12	13

Execution time. Several experiments were conducted to analyze the computational time required to obtain good solutions. We found that 1 minute per solution in the multi-start G+NO algorithm is enough time to achieve the greatest possible improvement of the solution obtained from the constructive phase. Figure 8.5 shows three 1-minute runs of the real instance, the best solution in each run being obtained in 13.6, 36.5, and 22.58 seconds.

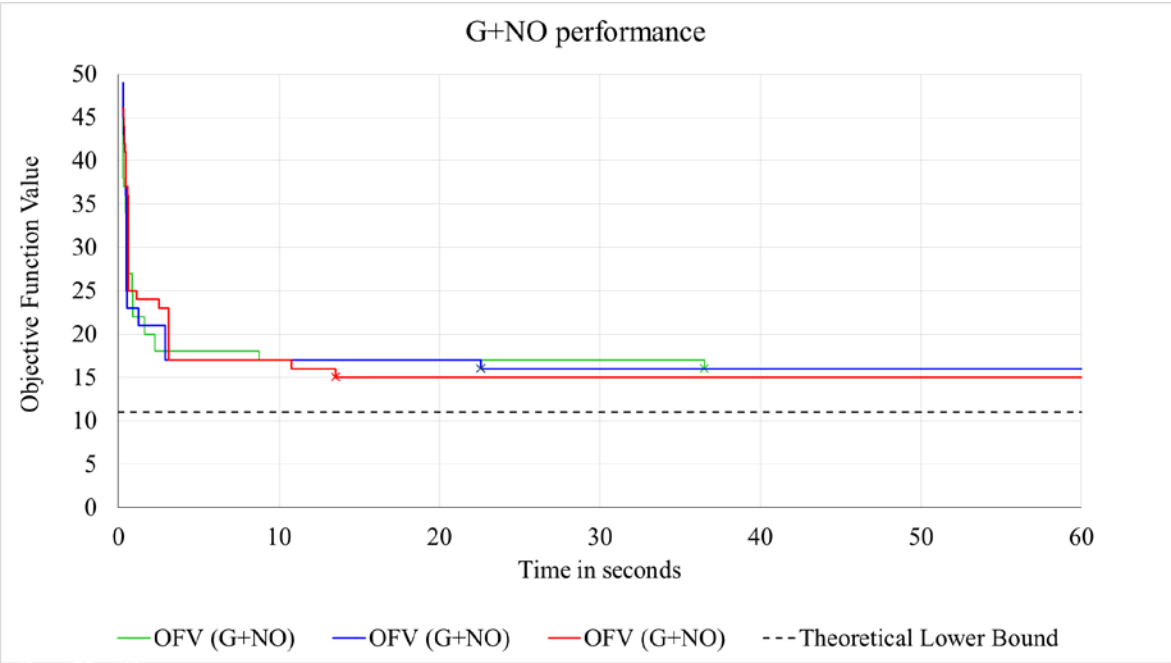


Figure 8.5. Three examples of 1 minute run of the G+NO algorithm for the real instance.

Figure 8.6 shows the G+NO performance for the most difficult problem, that is, scheduling shifts for 20 physicians, as used in the previous analysis. The upper graphs show three 1 minute G+NO runs of the instance, which obtains their best values in 18.191, 32.264, and 56.83

seconds. The second graph is a zoom of the previous graph, showing the points at which feasibility is recovered, the solutions achieve feasibility in 2.359, 1.512, and 3.641 seconds. The lower graph shows the G+NO run that provided the best solution for that instance in isolation. It reaches feasibility in 2.359 seconds and its best solution in 18.191 seconds, which is a value of 12 (the theoretical solution is 9, and the minimum solution provided by CPLEX in an hour is 23).

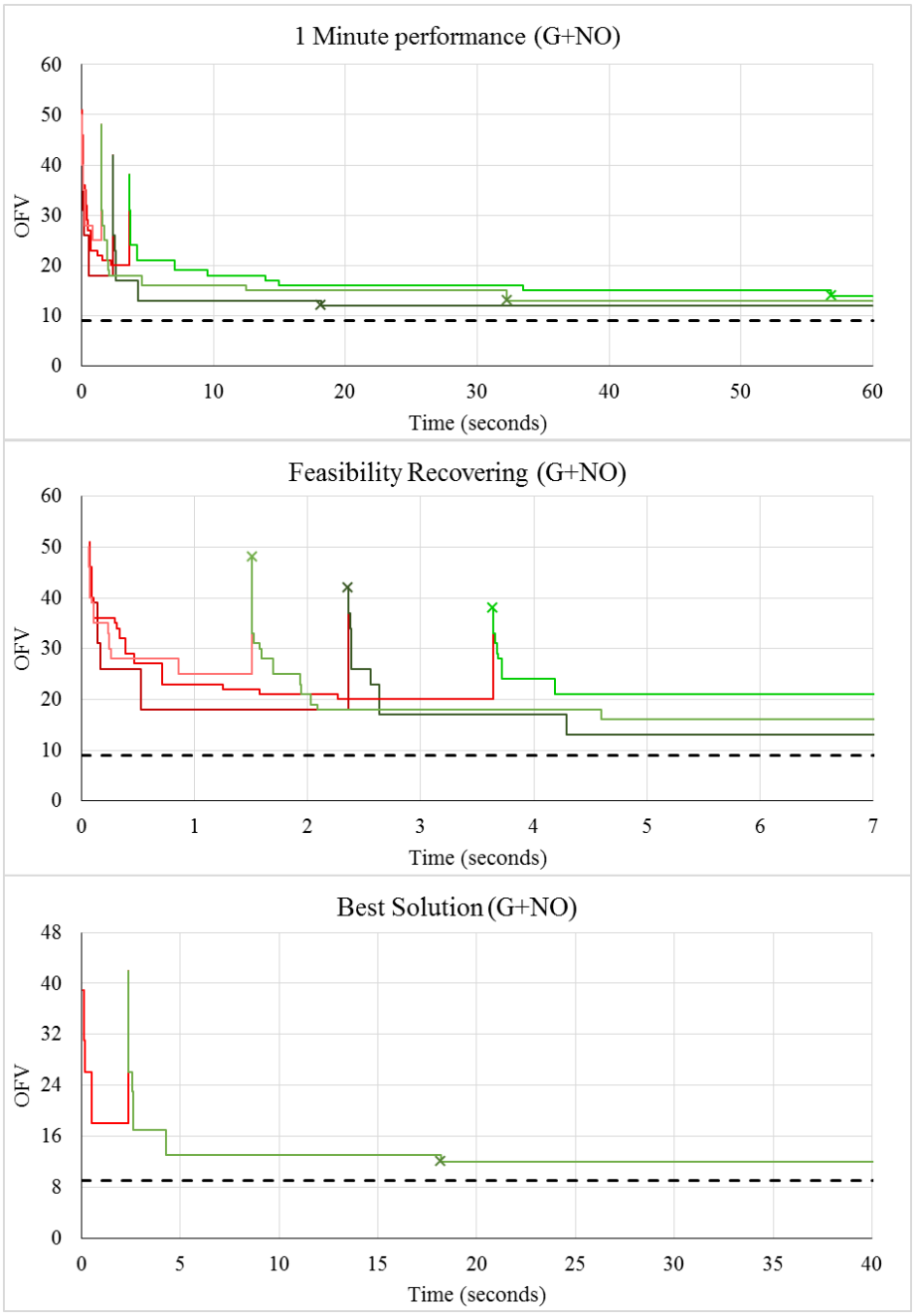


Figure 8.6. G+NO performance for the most difficult problem.

8.4 Implementation

The development of the mathematical models and algorithms underlying a rostering tool will involve the following last phase: specification and development of a reporting tool that displays solutions and provides performance reports [222].

Technical research, which regards similar problems to the one we address in this part of the thesis, usually focuses on the mathematical model and neglects the real life implications by making simplifications that were needed for the model to perform well. Moreover, in [211], [215]'s reviews have found that the number of research papers presenting a model that has been applied to a real life problem is very limited.

Fortunately, [215]'s review states that there is a trend towards an increasing willingness of hospitals to provide data and to conduct experimental studies and so most of the papers consider real-life data, but they do not implement their algorithms. In total, [215] reports that one third of the papers discusses the application of the theoretical model only in a test environment. Sometimes it is due to the lack of integration regarding medical staff, ED managers, and problem characteristics. Furthermore the software in which the algorithm is programmed hinder the implementation if it is commercial or not allowed by the ED's software system, which also entails a very costly and time-consuming endeavor.

However, the application of served theoretical insights in real life settings is an important part of research. It creates a feedback loop between practitioners and scientists by helping to test the practicability and validity of the theoretical models and employed modeling assumptions. It also helps scientist to test implementability of suggested process modifications to improve provided solutions and algorithms. Moreover, this increases medical staff's willingness to cooperate and enables researchers to evaluate the effect of their approach in their daily lives.

In this case study, the algorithm has been programmed in Java language (version 8), which is very useful to develop high performance portable applications for the widest range of possible computer platforms. We focused on creating graphical user interfaces and detailed reporting tools to allow them to evaluate the solutions and improve their acceptance of the new scheduling methods (see Figure 8.7 and Figure 8.8 that shows software screenshots).

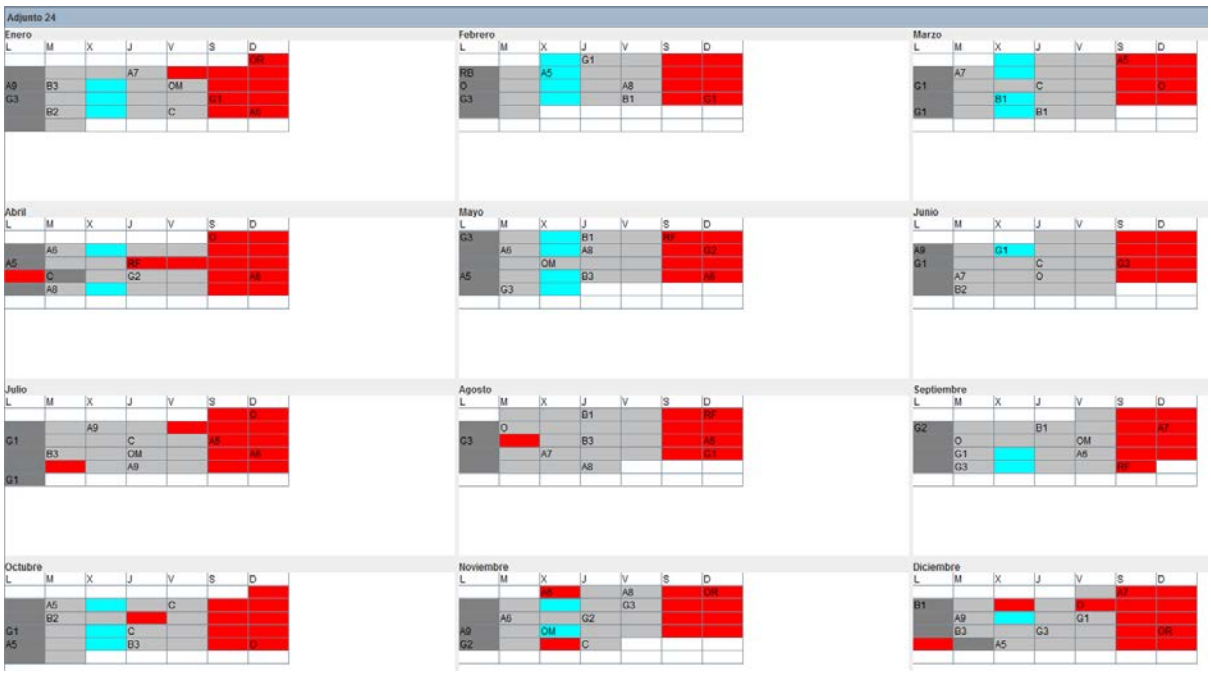


Figure 8.7. Screenshot of the software displaying the performance reports of the complete provided solution. Each column represent some objective defined by medical staff and each row represent a physician.

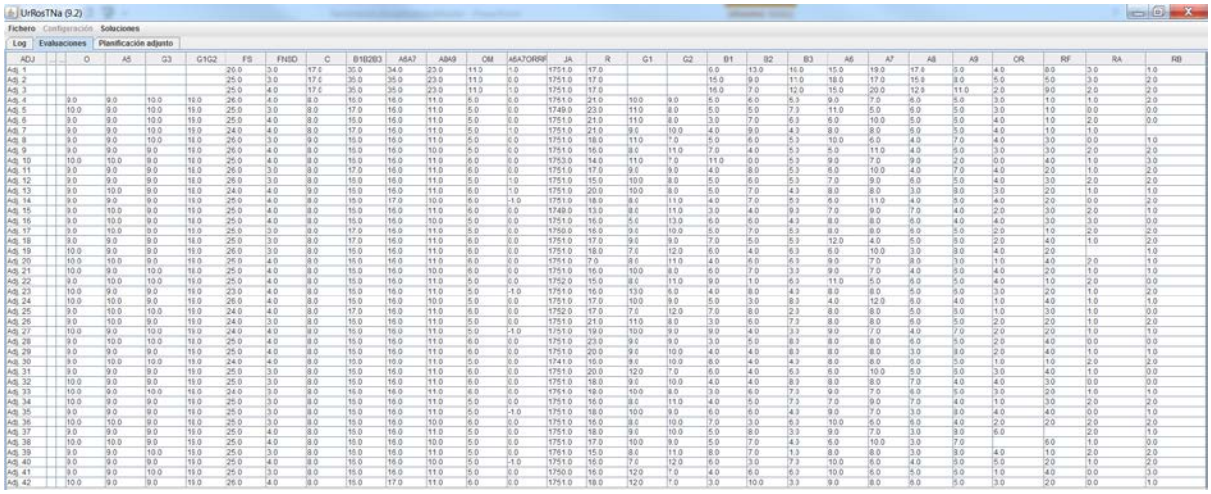


Figure 8.8. Screenshot of the software displaying the annual shifts assigned to “physician 24” in the provided solution (top left corner).

8.5 Conclusion

In this and the previous chapters, we have developed a new hybrid algorithm for solving the physician scheduling problem. One of the main features of the algorithm is that the number of shifts of each type that must be worked by each physician over the whole planning horizon is used to define the fitness function which determines his/her probability of being selected from the LoC. In this way, the construction step creates balanced solutions. In addition, the algorithm

also prioritizes the construction of feasible solutions by including in the LoC physicians who can feasibly work the shift being assigned in that step. As a result, the constructive phase usually obtains good quality solutions, because even infeasible solutions failing in only a few constraints can generally be repaired in the local search step. This step works by combining a shift-transfer process to reduce the number of infeasibilities, with a NFO process, to create new solutions to continue the search for shift-transfer chains. Once feasibility is achieved, the procedure continues with the NFO process alone in order to improve the balance of the solution.

The results show a clear superiority over ILP for realistically-sized instances; better results being achieved in a few minutes, as opposed to the 168 hours (an entire week) taken by CPLEX when real instances are solved. The resolution time, which can be up to several minutes in relatively large, heavily constrained problems, with little slack for physicians' working hours, can be considered satisfactory for use in practice. The algorithm can be applied for solving any scheduling problem that fits the general mathematical model presented in Section 6.3. It can handle different types of physicians and different types of shifts, with different types of constraints for each pairing (physician type, shift type). Thus, this general framework can fit other contexts, such as the scheduling of physicians in other health departments or police and fire department staff. In fact, the initial motivation of this research was the design of a general physician scheduling algorithm for any hospital department; the ER being the first department for which it was tested.

This study treats ergonomic constraints as hard constraints, although some could also be treated as soft constraints by penalizing any deviation beyond the bounds of the objective function. In this case, weights could be used in the objective function to prioritize the different objectives relative to each other and to other balancing criteria. This extension is quite common and straightforward to apply.

The use of NFO models to search large neighborhoods is one of the main features of this methodology. The use of exact methods to solve the network guarantees good, computationally economic, solution improvements, given the small size of the network (there are fewer nodes than physicians). Furthermore, the randomly constructed network favors the repeated use of this improvement step. It is worth mentioning that in the real problem, a narrow range of feasible schedules is obtained for annual hours worked (only two hours in the real case, with a window width of less than 0.05% of the average hours worked, 1,751), while the best solutions obtained by the scheduler at the hospital always provide ranges of more than 20 hours. Nevertheless, modeling with networks is a rich field that can be exploited to improve the procedure presented here. For example, currently, the costs do not discriminate between arcs, but they could express preferences to balance certain types of shifts or certain types of physicians. The algorithm is designed to build schedules from scratch but in order for it to be completely useful in practice, it should also be able to repair solutions. In this case, it would also be used for staff management purposes or for a minimal rearrangement of shifts when a physician is unable to attend work for any reason. However, this is a different problem, which,

while requiring its own formulation and solution procedures, it can usefully draw on the ideas used to develop the G+NO algorithm. This is a current topic of research.

Chapter 9 Conclusion

The research carried out in this thesis contributes to improvements in hospital Emergency Department (ED) management. The problems it addresses, which were raised by physicians working at the HCN ED and members of the q-UPHS research group, reflect the challenges currently facing ED managers. Hence, each area of research addressed in this thesis relates to the analysis of a real problem requiring a clear understanding of managers' objectives and real-life dynamics and complexity. In this practical context, mathematical modelling becomes particularly relevant, and the formulated model needs to capture all the key features of such complexity, while allowing for their variation and evolution over time. Consequently, simulation was selected as the mathematical tool to model the variability and stochasticity of the ED environment. The resulting simulation model considers the seasonality of patient arrival patterns, differentiated by severity, and mimics patient pathways through the ED, while reflecting the resource (including medical staff) consumption required for treatment. This mathematical model, presented in Chapter 2, overcomes some of the shortcomings of oversimplified queuing theory models and captures some important issues that previous simulation models have overlooked. With the help of a 3D animation, physicians were able to validate the simulation model for use in analyses aimed at improving patient-flow management. Thus, the first specific objective of the thesis, to "*Propose a quantitative framework (based on simulation models and their combination with optimization procedures) for the analysis of the problems involved in the dimensioning and assessment of management policies in hospital emergency services*" was fulfilled.

The first problem to be addressed by means of the simulation model in this research was the allocation of patients to physicians after triage. This involved developing new allocation rules, which proved to outperform the cyclic rule, used in some EDs because of its apparent fairness. The superiority of the new rules is demonstrated, first, by using the simulation model, as described in Chapter 4. The new allocation rules also take into account a factor usually neglected by patient-flow management policies: namely, workload stress in physicians. The inclusion of this factor in the assessment of patient management rules was motivated by a problem signaled by the physicians; namely, significant differences in the amount of work pending for each physician as the workshift advances. These differences, which were due to the differing treatment needs of their allocated patients, caused recurrent peaks of stress when there was a long queue of patients waiting to be seen (especially for the initial consultation).

The lack of a proper method of real-time physician stress measurement led to the research carried out in Chapter 3, which was aimed at developing a stress measurement method that could be used both as a KPI to assess the performance of patient-flow management policies and as a criterion for designing new ones. With this research, the thesis fulfilled the specific objective of "*developing a methodology for the real-time assessment of pending workload stress in physicians*". One of the proposed rules was also implemented in practice. Its successful implementation, from initial concept to practical application in the hospital, is illustrated in Chapter 4, where it is shown, through the analysis of real data, to outperform the current cyclic allocation rule. This research, together with that reported in Chapter 3, achieves the goal expressed in specific objective C of this thesis: "to *provide new patient-to-physician allocation methods with criteria including the workload and stress balancing across physicians and patient service quality"*.

The second step in ED patient-flow management is to determine the next patient to be seen once the physician has completed the preceding consultation. Decisions must depend on the severity and treatment phase of the pending patients. HCN physicians currently manage their own pending-patient portfolios, and their strategies vary. This problem is analyzed in Chapter 5, where it is modeled as a queuing network with priorities and patient re-entry. The simulation model enables the analysis of a new type of queuing discipline known as Accumulating Priority Queuing with the aim of obtaining the optimal solution (by means of simulation-based optimization methodology) and comparing it with those yielded by the pure priority discipline followed by physicians in practice. The studied mathematical model allows for access-time constraints, patient re-entry and different objective functions. These three aspects have not been considered simultaneously in a realistic ED environment in any previous related research, either from the queuing theory or the simulation modelling perspective. Therefore, the research described in this Chapter 5 accomplishes specific objective D: "to a*nalyze alternatives to pure priority rules for managing the queue of patients awaiting initial emergency assessment by a physician or reevaluation following tests and/or diagnosis"*.

The second part of the thesis addresses the physician scheduling problem, which is a combinatorial optimization problem posing particular difficulty when considering all the constraints and objectives observed in practice. Chapter 6 models these constraints and objectives using mathematical programming, while Chapter 7 exposes the new problem-solving heuristic. A key feature of this algorithm is the greedy constructive phase, which is guided by solving a linear problem in combination with a simple memory structure. Initial good solutions are very quickly obtained, but they can be unfeasible in heavily constrained cases. The subsequent improvement phase combines a repair strategy based on variable neighborhood search with network optimization. As far as can be ascertained, this is the first proposal for such a strategy. A computational analysis and a real-case solution, presented in Chapter 8, demonstrate the quality of the solutions and the good behavior of the methodology. The real-case solution is also used in practice to program a year-long workshift schedule for the 42 physicians working at the HCN. The research described in chapters 6, 7 and 8 therefore enables

the satisfactory fulfillment of objective E of this thesis: "*to design efficient algorithms for solving the physician workshift assignment problem taking into account all real ergonomic constraints while balancing the workload*".

Thus, the research carried out in this thesis accomplishes the general objective which is "*to develop methodologies and algorithms enabling operational and tactical decisions to improve hospital emergency services, both from the patient and the workforce perspective*". However, these problems are not completely solved and new research opportunities and challenges are revealed by the results presented.

Chapter 4 proposes several patient–to-physician allocation methods, but only one, the simplest in terms of the information required for its implementation, is fully analyzed. Our purpose is to perform a thorough analysis of these policies for optimizing patient waiting times, reducing stress in doctors and equitably distributing the workload. These three criteria may conflict with one another; especially in a context of different doctor service rates, a situation not considered in this thesis. The modeling of individual physician behavior (different work speeds, different resilience to stress, etc.) also reveals the need to extend the discrete event simulation model and combine it with an agent-based simulation model.

The analysis of patient queue management policies has focused on pure priority disciplines, which only need to consider the severity of each patient, and APQ disciplines, which also take into account the time spent waiting by the first patient of each type. Although many EDs currently incorporate patient geo-positioning, such data are not taken into account in management policy design. Our aim is to investigate new management policies incorporating patient severity, treatment phase, gps, length of stay and time spent waiting by all patients in the ED, also including the probability that new arrivals are not seen within the target access time. The new policies will address continuous healthcare improvement, minimization of LoS, overcrowding, and comply with target access times.

Interaction with the ED physician schedule manager suggests new heuristic approaches to solving the scheduling problem. The idea is to develop new algorithms incorporating the different objectives and rationalizing their importance. In the case of the greedy algorithm, for example, this means creating the solution not by following a sequence of steps, as in the current manner, but by achieving an ordered set of goals. This strategy allows for the reiterative use of a greedy algorithm in the initial stages of emergency care when new shift assignments are not highly constrained, leaving the use of small linear programming problems for the final stages, when new assignments are heavily constrained and good feasible ones are hard to find. Furthermore, new models and algorithms are required when, instead of balancing the workload, the aim is to satisfy the wishes of the physician.

References

[1] P. Tudela and J. M. Mòdol, “On hospital emergency department crowding,” *Emergencias Rev. la Soc. Esp. Med. Emergencias*, vol. 27, no. 2, pp. 113–120, 2015.

[2] C. Narvarro García, “La utlización de los servicios de urgencia y la Tragedia de los Comunes,” *Tesis Dr.*, no. Departamento de Economía Aplicada y Gestión Pública. Facultad de ciencias económicas y empresariales, UNED, p. 166, 2015.

[3] G. Hardin, H. Bonfil-Sánchez, and (Traductor), “La tragedia de los comunes (The tragedy of commons),” *Gac. Ecológica*, no. 37, 1995.

[4] Defensor del Pueblo (Ombudsman), “Las urgencias hospitalarias en el Sistema Nacional de Salud: derechos y garantías de los pacientes.”

[5] M. Gendreau *et al.*, “Physician Scheduling in Emergency Rooms,” in *Practice and Theory of Automated Timetabling VI*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 53–66.

[6] M. A. Arostegui, S. N. Kadipasaoglu, and B. M. Khumawala, “An empirical comparison of Tabu Search, Simulated Annealing, and Genetic Algorithms for facilities location problems,” *Int. J. Prod. Econ.*, vol. 103, no. 2, pp. 742–754, Oct. 2006.

[7] M. W. Carter and S. D. Lapierre, “Scheduling Emergency Room Physicians,” *Health Care Manag. Sci.*, vol. 4, no. 4, pp. 347–360, 2001.

[8] X. Wang, “Emergency Department Staffing: A Separated Continuous Linear Programming Approach,” *Math. Probl. Eng.*, vol. 2013, pp. 1–8, 2013.

[9] L. C and S. Appa Iyer, “Application of queueing theory in health care: A literature

review," *Oper. Res. Heal. Care*, vol. 2, no. 1–2, pp. 25–39, Mar. 2013.

[10] C. Azcarate, L. Esparza, and F. Mallor, "The problem of the last bed: Contextualization and a new simulation framework for analyzing physician decisions," *Omega*, p. 102120, Sep. 2019.

[11] A. Kolker, "Process modeling of emergency department patient flow: Effect of patient length of stay on ED diversion," *J. Med. Syst.*, vol. 32, no. 5, pp. 389–401, 2008.

[12] S. C. Brailsford, P. R. Harper, B. Patel, and M. Pitt, "An analysis of the academic literature on simulation and modelling in health care," *J. Simul.*, vol. 3, no. 3, pp. 130–140, Sep. 2009.

[13] M. M. Günal and M. Pidd, "Discrete event simulation for performance modelling in health care: A review of the literature," *J. Simul.*, vol. 4, no. 1, pp. 42–51, 2010.

[14] L. Vanbrabant, K. Braekers, K. Ramaekers, and I. Van Nieuwenhuyse, "Simulation of emergency department operations: A comprehensive review of KPIs and operational improvements," *Comput. Ind. Eng.*, vol. 131, no. March, pp. 356–381, May 2019.

[15] S. Saghafian, G. Austin, and S. J. Traub, "Operations research/management contributions to emergency department patient flow optimization: Review and research prospects," *IIE Trans. Healthc. Syst. Eng.*, vol. 5, no. 2, pp. 101–123, Apr. 2015.

[16] S. Saghafian, G. Austin, and S. J. Traub, "Operations research/management contributions to emergency department patient flow optimization: Review and research prospects," *IIE Trans. Healthc. Syst. Eng.*, vol. 5, no. 2, pp. 101–123, Apr. 2015.

[17] M. Gul and A. F. Guneri, "A comprehensive review of emergency department simulation applications for normal and disaster conditions," *Comput. Ind. Eng.*, vol. 83, no. 8, pp. 327–344, May 2015.

[18] S. Robinson, "Conceptual modelling for simulation Part I: definition and requirements," *J. Oper. Res. Soc.*, vol. 59, no. 3, pp. 278–290, Mar. 2008.

[19] M. M. Gunal, "A guide for building hospital simulation models," *Heal. Syst.*, vol. 1, no. 1, pp. 17–25, Jun. 2012.

[20] A. Kumar and R. Kapur, “Discrete simulation application-scheduling staff for the emergency room,” *Proc. 21st Conf. Winter*, 1989.

[21] M. D. Rossetti, G. F. Trzcinski, and S. A. Syverud, “Emergency department simulation and determination of optimal attending physician staffing schedules,” in *Winter Simulation Conference Proceedings*, 1999, vol. 2, no. 1532, pp. 1532–1540.

[22] Y.-H. Kuo, O. Rado, B. Lupia, J. M. Y. Leung, and C. A. Graham, “Improving the efficiency of a hospital emergency department: a simulation study with indirectly imputed service-time distributions,” *Flex. Serv. Manuf. J.*, vol. 28, no. 1–2, pp. 120–147, Jun. 2016.

[23] S. Kelton, Sadowski, *Simulation with Arena Fourth Edition*, vol. 3. McGraw-Hill, 2002.

[24] E. Cinlar, *Introduction to Stochastic Processes*. 1975.

[25] L. M. Leemis, “Nonparametric Estimation of the Cumulative Intensity Function for a Nonhomogeneous Poisson Process,” *Manage. Sci.*, vol. 37, no. 7, pp. 886–900, Jul. 1991.

[26] S. G. Eick, W. A. Massey, and W. Whitt, “The Physics of the Mt/G/∞ Queue,” *Oper. Res.*, vol. 41, no. 4, pp. 731–742, Aug. 1993.

[27] M. E. H. Ong *et al.*, “Using demand analysis and system status management for predicting ED attendances and rostering,” *Am. J. Emerg. Med.*, vol. 27, no. 1, pp. 16–22, 2009.

[28] S. Peiró, J. Librero, M. Ridao, and E. Bernal-Delgado, “Variabilidad en la utilización de los servicios de urgencias hospitalarios del Sistema Nacional de Salud,” *Gac. Sanit.*, vol. 24, no. 1, pp. 6–12, Jan. 2010.

[29] F. Aguado-Correa, M. Herrera-Carranza, and N. Padilla-Garrido, “Variability and Overcrowding Management: Ongoing Challenge for Spanish Hospital Emergency Departments,” *J. Health Manag.*, vol. 18, no. 2, pp. 218–230, 2016.

[30] M. I. Cano del Pozo *et al.*, “Estudio de la frecuentación de un servicio de urgencias extrahospitalario,” *Emergencias*, vol. 20, no. 3, pp. 179–186, 2008.

[31] Z. Zeng, X. Ma, Y. Hu, J. Li, and D. Bryant, "A Simulation Study to Improve Quality of Care in the Emergency Department of a Community Hospital," *J. Emerg. Nurs.*, vol. 38, no. 4, pp. 322–328, Jul. 2012.

[32] M. A. Ahmed and T. M. Alkhamis, "Simulation optimization for an emergency department healthcare unit in Kuwait," *Eur. J. Oper. Res.*, vol. 198, no. 3, pp. 936–942, Nov. 2009.

[33] M. Gunal and M. Pidd, "Understanding Accident and Emergency Department Performance using Simulation," in *Proceedings of the 2006 Winter Simulation Conference*, 2006, pp. 446–452.

[34] L. B. Holm and M. Barra, "The consequences of how subject matter expert estimates are interpreted and modelled, demonstrated by an emergency department des model comparing triangular and beta distributions," in *Proceedings of the 2011 Winter Simulation Conference (WSC)*, 2011, pp. 3649–3656.

[35] A. Gupta, "Simulation modeling and analysis," in *Decision Sciences: Theory and Practice*, McGraw-Hill Education, 2016, pp. 801–886.

[36] D. H. Uyeno, C. Seeberg, and H. Dean, "practical methodology for ambulance location Faculty," pp. 79–87, 1971.

[37] P. M. S. Silva and L. R. Pinto, "Emergency medical systems analysis by simulation and optimization," in *Proceedings of the 2010 Winter Simulation Conference*, 2010, no. 1989, pp. 2422–2432.

[38] O. Balci, "Verification validation and accreditation of simulation models," in *Proceedings of the 29th conference on Winter simulation - WSC '97*, 1997, pp. 135–141.

[39] J. P. C. Kleijnen, "Verification and validation of simulation models," no. mimic, 1995.

[40] J. Banks, *Discrete-event system simulation*. Prentice Hall, 2010.

[41] C. Baril, V. Gascon, and J. Miller, "Design of experiments and discrete-event simulation to study oncology nurse workload," *IISE Trans. Healthc. Syst. Eng.*, vol. 0, no. 0, pp. 1–13, Oct. 2019.

[42] L. Aboueljinane, E. Sahin, and Z. Jemai, “A review on simulation models applied to emergency medical service operations,” *Comput. Ind. Eng.*, vol. 66, no. 4, pp. 734–750, Dec. 2013.

[43] A. van Gestel, J. L. Severens, C. A. B. Webers, H. J. M. Beckers, N. M. Jansonius, and J. S. A. G. Schouten, “Modeling Complex Treatment Strategies: Construction and Validation of a Discrete Event Simulation Model for Glaucoma,” *Value Heal.*, vol. 13, no. 4, pp. 358–367, Jun. 2010.

[44] M. Reynolds *et al.*, “Using discrete event simulation to design a more efficient hospital pharmacy for outpatients,” *Health Care Manag. Sci.*, vol. 14, no. 3, pp. 223–236, Sep. 2011.

[45] S. G. Henderson and A. J. Mason, “Ambulance Service Planning: Simulation and Data Visualisation,” in *Operations Research and Health Care*, Boston: Kluwer Academic Publishers, 2006, pp. 77–102.

[46] A. Ingolfsson, E. Erkut, and S. Budge, “Simulation of single start station for Edmonton EMS,” *J. Oper. Res. Soc.*, vol. 54, no. 7, pp. 736–746, 2003.

[47] L. Aboueljinane, Z. Jemai, and E. Sahin, “Reducing ambulance response time using simulation: The case of Val-de-Marne department Emergency Medical service,” in *Proceedings Title: Proceedings of the 2012 Winter Simulation Conference (WSC)*, 2012, pp. 1–12.

[48] M. Lubicz and B. Mielczarek, “Simulation modelling of emergency medical services,” *Eur. J. Oper. Res.*, vol. 29, no. 2, pp. 178–185, 1987.

[49] W. Luo, J. Cao, M. Gallagher, and J. Wiles, “Estimating the intensity of ward admission and its effect on emergency department access block,” *Stat. Med.*, vol. 32, no. 15, pp. 2681–2694, Jul. 2013.

[50] M. Twomey, L. A. Wallis, and J. E. Myers, “Limitations in validating emergency department triage scales,” *Emerg. Med. J.*, vol. 24, no. 7, pp. 477–479, Jul. 2007.

[51] I. Rockwell Automation Technologies, “Arena.” 2016.

[52] "q-UPHS: quantitative methods for Uplifting the Performance of Health Services." [Online]. Available: http://www.unavarra.es/digitalAssets/233/233798_100000proyecto05_video01.mp4.

[53] J. D. Schuur and A. K. Venkatesh, "The Growing Role of Emergency Departments in Hospital Admissions," *N. Engl. J. Med.*, vol. 367, no. 5, pp. 389–391, 2012.

[54] NHS England, "Reducing Emergency Admissions," 2018.

[55] N. Tang, J. Stein, R. Y. Hsia, J. H. Maselli, and R. Gonzales, "Trends and Characteristics of US Emergency Department Visits, 1997-2007," *JAMA*, vol. 304, no. 6, p. 664, Aug. 2010.

[56] B. C. Strunk, P. B. Ginsburg, and M. I. Banker, "The effect of population aging on future hospital demand," *Health Aff.*, vol. 25, no. 3, 2006.

[57] National Center for Health Statistics, "Health, United States, 2016: With chartbook on long-term trends in health," *Cent. Dis. Control*, pp. 314–317, 2017.

[58] M. McHugh, K. Van Dyke, M. McClelland, and D. Moss, "Improving Patient Flow and Reducing Emergency Department Crowding: A Guide for Hospitals," Rockville, MD, 2011.

[59] J. L. Wiler, R. T. Griffey, and T. Olsen, "Review of Modeling Approaches for Emergency Department Patient Flow and Crowding Research," *Acad. Emerg. Med.*, vol. 18, no. 12, pp. 1371–1379, Dec. 2011.

[60] Y. Ding, E. Park, M. Nagarajan, and E. Grafstein, "Patient Prioritization in Emergency Department Triage Systems: An Empirical Study of Canadian Triage and Acuity Scale (CTAS)," *SSRN Electron. J.*, 2018.

[61] S. J. Traub *et al.*, "Emergency department rapid medical assessment: overall effect and mechanistic considerations.," *J. Emerg. Med.*, vol. 48, no. 5, pp. 620–7, May 2015.

[62] M. F. Kamali, T. Tezcan, and O. Yildiz, "When to use provider triage in emergency departments," *Manage. Sci.*, vol. 65, no. 3, pp. 1003–1019, Mar. 2019.

[63] J. H. Han, C. Zhou, D. J. France, S. Zhong, and I. Jones, “The Effect of Emergency Department Expansion on Emergency Department Overcrowding,” pp. 338–343, 2007.

[64] S. Russ, I. Jones, D. Aronsky, R. S. Dittus, and C. M. Slovis, “Placing Physician Orders at Triage: The Effect on Length of Stay,” *Ann. Emerg. Med.*, vol. 56, no. 1, pp. 27–33, Jul. 2010.

[65] S. J. Traub, A. C. Bartley, V. D. Smith, R. Didehban, C. A. Lipinski, and S. Saghafian, “Physician in Triage Versus Rotational Patient Assignment,” *J. Emerg. Med.*, vol. 50, no. 5, pp. 784–790, May 2016.

[66] S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick, “Complexity-Augmented Triage: A Tool for Improving Patient Safety and Operational Efficiency,” *Manuf. Serv. Oper. Manag.*, vol. 16, no. 3, pp. 329–345, Jul. 2014.

[67] S. Ieraci, E. Digiusto, P. Sonntag, L. Dann, and D. Fox, “Streaming by case complexity: Evaluation of a model for emergency department Fast Track,” *EMA - Emerg. Med. Australas.*, vol. 20, no. 3, pp. 241–249, 2008.

[68] D. I. Ben-Tovim *et al.*, “Redesigning care at the Flinders Medical Centre: clinical process redesign using ‘lean thinking’.,” *Med. J. Aust.*, vol. 188, no. 6 Suppl, pp. 27–31, 2008.

[69] D. L. King, D. I. Ben-Tovim, and J. Bassham, “Redesigning emergency department patient flows: Application of Lean Thinking to health care,” *EMA - Emerg. Med. Australas.*, vol. 18, no. 4, pp. 391–397, 2006.

[70] B. S. J. Welch, “Patient Segmentation: Redesigning Flow,” 2009.

[71] S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick, “Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments,” *Oper. Res.*, vol. 60, no. 5, pp. 1080–1097, Oct. 2012.

[72] S. Welch, J. Augustine, C. A. Camargo, and C. Reese, “Emergency Department Performance Measures and Benchmarking Summit,” *Acad. Emerg. Med.*, vol. 13, no. 10, pp. 1074–1080, Oct. 2006.

[73] S. J. Welch, B. R. Asplin, S. Stone-Griffith, S. J. Davidson, J. Augustine, and J. Schuur, "Emergency Department Operational Metrics, Measures and Definitions: Results of the Second Performance Measures and Benchmarking Summit," *Ann. Emerg. Med.*, vol. 58, no. 1, pp. 33–40, Jul. 2011.

[74] The National Institute for Occupational Safety and Health (NIOSH), "Stress...At Work," 01-Jan-1999. [Online]. Available: https://www.cdc.gov/niosh/docs/99-101/. [Accessed: 03-Jan-2019].

[75] S. Cazabat, B. Barthe, and N. Cascino, "Work load and job stress: two facets of the same situation? Exploratory study in a gerontology department," *Perspect. Interdiscip. sur le Trav. la santé*, no. 10–1, pp. 0–19, 2008.

[76] Department of Health and Human Services, "Exposure to Stress. Occupational Hazards in Hospitals.," 2008.

[77] M. Estryn-Behar *et al.*, "Emergency physicians accumulate more stress factors than other physicians-results from the French SESMAT study," *Emerg. Med. J.*, vol. 28, no. 5, pp. 397–410, May 2011.

[78] G. Matthews, "Multidimensional Profiling of Task Stress States for Human Factors," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 58, no. 6, pp. 801–813, Sep. 2016.

[79] M. Pottier and M. Estryn-Behar, "L'ergonomie du travail infirmier," *Trav. Hum.*, vol. 43, no. 1, pp. 121–153, 1980.

[80] C. Martin and C. Gadbois, "36. L'ergonomie à l'hôpital," in *Ergonomie*, Ed P. Falz., PUF, Paris.: Presses Universitaires de France, 2004, p. 603.

[81] M. Estryn-Behar and J.-P. Fouillot, "Etude de a charge physique du personnel soignant," *Doc. pour le Médecin du Trav.*, 1990.

[82] C. Nogareda Cuixart, "NTP 275:Carga mental en el trabajo hospitalario: Guía para su valoración," 1991.

[83] L. Phipps, "Stress among doctors and nurses in the emergency department of a general hospital.," *C. Can. Med. Assoc. J.*, vol. 139, no. 5, pp. 375–6, Sep. 1988.

[84] K. L. Keller and W. J. Koenig, “Sources of stress and satisfaction in emergency practice,” *J. Emerg. Med.*, vol. 7, no. 3, pp. 293–299, May 1989.

[85] W. W. Lambert and R. S. Lazarus, “Psychological Stress and the Coping Process,” *Am. J. Psychol.*, vol. 83, no. 4, p. 634, 1970.

[86] G. P. Chrousos and P. W. Gold, “The Concepts of Stress and Stress System Disorders,” *JAMA*, vol. 267, no. 9, p. 1244, Mar. 1992.

[87] H. Selye, *The stress of life*. New York: McGraw-Hill, 1956.

[88] R. S. Lazarus, *Emotion and Adaptation*. New York, NY, US: Oxford University Press, 1991.

[89] R. S. Lazarus and S. Folkman, *Stress, appraisal, and coping*. Springer publishing company, 1984.

[90] J. E. McGrath, “Methodological problems in research on stress.,” *Ser. Clin. Community Psychol. Achiev. Stress. Anxiety*, pp. 19–48, 1982.

[91] S. G. Hill, H. P. Lavecchia, J. C. Byers, A. C. Bittner, A. L. Zaklad, and R. E. Christ, “Comparison of four subjective workload rating scales,” *Hum. Factors*, vol. 34, no. 4, pp. 429–439, 1992.

[92] S. Levin *et al.*, “Tracking Workload in the Emergency Department,” *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 48, no. 3, pp. 526–539, Sep. 2006.

[93] K. Brodman, A. J. Erdmann, I. Lorge, H. G. Wolff, and T. H. Broadbent, “THE CORNELL MEDICAL INDEX,” *J. Am. Med. Assoc.*, vol. 140, no. 6, p. 530, Jun. 1949.

[94] C. Maslach and S. Jackson, *Maslach burnout inventory: research edition; manual*. Consulting Psychologists Press, 1981.

[95] G. Matthews *et al.*, “Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry.,” *Emotion*, vol. 2, no. 4, pp. 315–340, Dec. 2002.

[96] R. R. Holt, “Occupational stress.,” in *Handbook of stress: Theoretical and clinical*

aspects, 2nd ed., New York, NY, US: Free Press, 1993, pp. 342–367.

[97] A. Ferjani, A. Ammar, H. Pierreval, and S. Elkosantini, “A simulation-optimization based heuristic for the online assignment of multi-skilled workers subjected to fatigue in manufacturing systems,” *Comput. Ind. Eng.*, vol. 112, pp. 663–674, 2017.

[98] J. Razmi, H. Rafiei, and M. Hashemi, “Designing a decision support system to evaluate and select suppliers using fuzzy analytic network process,” *Comput. Ind. Eng.*, vol. 57, no. 4, pp. 1282–1290, Nov. 2009.

[99] A. Cossari, J. C. Ho, G. Paletta, and A. J. Ruiz-Torres, “A new heuristic for workload balancing on identical parallel machines and a statistical perspective on the workload balancing criteria,” *Comput. Oper. Res.*, vol. 39, no. 7, pp. 1382–1393, 2012.

[100] SAS Institute Inc., “JMP®, Version 13,.” SAS Institute Inc., Cary, NC,.

[101] R. K. Meyer and C. J. Nachtsheim, “The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs,” *Technometrics*, vol. 37, no. 1, pp. 60–69, Feb. 1995.

[102] A. Tversky and D. Kahneman, “Judgment under uncertainty: Heuristics and biases,” *Science (80-.).*, vol. 185, no. 4157, pp. 1124–1131, Jul. 1974.

[103] G. A. Miller, “The magical number seven, plus or minus two: some limits on our capacity for processing information.,” *Psychol. Rev.*, vol. 63, no. 2, pp. 81–97, 1956.

[104] M. Greller and C. K. Parsons, “Psychosomatic Complaints Scale of Stress: Measure Development and Psychometric Properties,” *Educ. Psychol. Meas.*, vol. 48, no. 4, pp. 1051–1065, Dec. 1988.

[105] S. M. Casner and B. F. Gore, “Measuring and evaluating workload: A primer,” *NASA Tech. Memo.*, vol. 216395, p. 2010, 2010.

[106] E. Svensson, “Different ranking approaches defining association and agreement measures of paired ordinal data,” *Stat. Med.*, vol. 31, no. 26, pp. 3104–3117, Nov. 2012.

[107] K. P. Nelson and D. Edwards, “Measures of agreement between many raters for ordinal

classifications," *Stat. Med.*, vol. 34, no. 23, pp. 3116–3132, Oct. 2015.

[108] D. Marasini, P. Quatto, and E. Ripamonti, "Assessing the inter-rater agreement for ordinal data through weighted indexes," *Stat. Methods Med. Res.*, vol. 25, no. 6, pp. 2611–2633, Dec. 2016.

[109] C. Roberts and R. McNamee, "Assessing the reliability of ordered categorical scales using kappa-type statistics," *Stat. Methods Med. Res.*, vol. 14, no. 5, pp. 493–514, Oct. 2005.

[110] J. C. Nelson and M. S. Pepe, "Statistical description of interrater variability in ordinal ratings," *Stat. Methods Med. Res.*, vol. 9, no. 5, pp. 475–496, Oct. 2000.

[111] E. L. Korn and B. I. Graubard, *Analysis of Health Surveys*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1999.

[112] B. J. Gajewski, S. Thompson, N. Dunton, A. Becker, and M. Wrona, "Inter-rater reliability of nursing home surveys: a Bayesian latent class approach," *Stat. Med.*, vol. 25, no. 2, pp. 325–344, Jan. 2006.

[113] H. C. Kraemer, "Measurement of reliability for categorical data in medical research," *Stat. Methods Med. Res.*, vol. 1, no. 2, pp. 183–199, Aug. 1992.

[114] F. M. Lord, M. R. Novick, and A. Birnbaum, *Statistical theories of mental test scores*. Information Age Pub, 1968.

[115] L. J. Cronbach, *The Dependability of behavioral measurements: theory of generalizability for scores and profiles*. Wiley, 1972.

[116] M. Shalon and M. J. Strube, "Type A behavior and emotional responses to uncertainty: A test of the self-appraisal model," *Motiv. Emot.*, vol. 12, no. 4, pp. 385–398, Dec. 1988.

[117] H. Rastegary and F. J. Landy, "The Interactions among Time Urgency, Uncertainty, and Time Pressure," in *Time Pressure and Stress in Human Judgment and Decision Making*, Boston, MA: Springer US, 1993, pp. 217–239.

[118] M. Linzer *et al.*, "Managed care, time pressure, and physician job satisfaction: Results

from the physician worklife study," *J. Gen. Intern. Med.*, vol. 15, no. 7, pp. 441–450, Jul. 2000.

[119] T. L. Saaty, "The Analytical Hierarchy Process." New York, NY., 1980.

[120] R. Handfield, S. V. Walton, R. Sroufe, and S. A. Melnyk, "Applying environmental criteria to supplier assessment: A study in the application of the Analytical Hierarchy Process," *Eur. J. Oper. Res.*, vol. 141, no. 1, pp. 70–87, 2002.

[121] T. L. Saaty, "How to make a decision: The analytic hierarchy process," *Eur. J. Oper. Res.*, vol. 48, no. 1, pp. 9–26, 1990.

[122] B. Shneiderman, "Handbook of Human Factors and Ergonomics (4th ed.)," *Int. J. Hum. Comput. Interact.*, vol. 28, no. 12, pp. 838–838, Dec. 2012.

[123] G. Matthews, J. Szalma, A. R. Panganiban, C. Neubauer, and J. S. Warm, "Profiling task stress with the dundee state questionnaire," in *Psychology of Stress: New Research*, vol. 1, no. February, 2013, pp. 49–91.

[124] D. P. Andrulis, A. Kellermann, E. A. Hintz, B. B. Hackman, and V. B. Weslowski, "Emergency departments and crowding in United States teaching hospitals," *Ann. Emerg. Med.*, vol. 20, no. 9, pp. 980–986, Sep. 1991.

[125] D. M. Fatovich, "Recent developments: Emergency medicine," *BMJ*, vol. 324, no. 7343, pp. 958–962, Apr. 2002.

[126] R. W. Derlet, J. R. Richards, and R. L. Kravitz, "Frequent Overcrowding in U.S. Emergency Departments," *Acad. Emerg. Med.*, vol. 8, no. 2, pp. 151–155, Feb. 2001.

[127] R. W. Derlet and J. R. Richards, "Emergency department overcrowding in Florida, New York, and Texas," *South. Med. J.*, vol. 95, no. 8, pp. 846–849, 2002.

[128] R. W. Derlet and J. R. Richards, "Overcrowding in the nation's emergency departments: Complex causes and disturbing effects," *Ann. Emerg. Med.*, vol. 35, no. 1, pp. 63–68, 2000.

[129] J. R. Richards, M. L. Navarro, and R. W. Derlet, "Survey of directors of emergency

departments in California on overcrowding," *West. J. Med.*, vol. 172, no. 6, pp. 385–388, 2000.

[130] R. W. Derlet, "Overcrowding in emergency departments: Increased demand and decreased capacity," *Ann. Emerg. Med.*, vol. 39, no. 4, pp. 430–432, 2002.

[131] S. Liu, "Impact of Critical Bed Status on Emergency Department Patient Flow and Overcrowding," *Acad. Emerg. Med.*, vol. 10, no. 4, pp. 382–385, Apr. 2003.

[132] B. R. Asplin, D. J. Magid, K. V Rhodes, L. I. Solberg, N. Lurie, and C. A. Camargo, "A conceptual model of emergency department crowding," *Ann. Emerg. Med.*, vol. 42, no. 2, pp. 173–180, Aug. 2003.

[133] B. Bursch, J. Beezy, and R. Shaw, "Emergency department satisfaction: What matters most?," *Ann. Emerg. Med.*, vol. 22, no. 3, pp. 586–591, Mar. 1993.

[134] A. Maitra and C. Chikhani, "Patient satisfaction in an urban accident and emergency department.," *Br. J. Clin. Pract.*, vol. 46, no. 3, pp. 182–4, 1992.

[135] D. DeBehnke and M. C. Decker, "The effects of a physician-nurse patient care team on patient satisfaction in an academic ED," *Am. J. Emerg. Med.*, vol. 20, no. 4, pp. 267–270, Jul. 2002.

[136] F. L. Lau and K. P. Leung, "Waiting time in an urban accident and emergency department--a way to improve it.," *J. Accid. Emerg. Med.*, vol. 14, no. 5, pp. 299–301; discussion 302-3, 1997.

[137] D. W. Spaite *et al.*, "Rapid process redesign in a university-based emergency department: Decreasing waiting time intervals and improving patient satisfaction," *Ann. Emerg. Med.*, vol. 39, no. 2, pp. 168–177, Feb. 2002.

[138] M. D. Fottler and R. C. Ford, "Managing patient waits in hospital emergency departments.," *Health Care Manag. (Frederick).*, vol. 21, no. 1, pp. 46–61, Sep. 2002.

[139] D. N. Kyriacou, V. Ricketts, P. L. Dyne, M. D. McCollough, and D. A. Talan, "A 5-year time study analysis of emergency department patient care efficiency," *Ann. Emerg. Med.*, vol. 34, no. 3, pp. 326–335, 1999.

[140] S. J. Traub *et al.*, "Emergency Department Rotational Patient Assignment," *Ann. Emerg. Med.*, vol. 67, no. 2, pp. 206–215, Feb. 2016.

[141] T. Ünlüyurt and Y. Tunçer, "Estimating the performance of emergency medical service location models via discrete event simulation," *Comput. Ind. Eng.*, vol. 102, pp. 467–475, Dec. 2016.

[142] D. R. Smith and W. Whitt, "Resource Sharing for Efficiency in Traffic Systems," *Bell Syst. Tech. J.*, vol. 60, no. 1, pp. 39–55, Jan. 1981.

[143] B. E, H. HL, J. A, and Erlang AK, *The life and works of A.K. Erlang*. 1948.

[144] G. D. Eppen, "Note—Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem," *Manage. Sci.*, vol. 25, no. 5, pp. 498–501, May 1979.

[145] L. Kleinrock, *QUEUEING SYSTEMS, Volume 2*, vol. 6. New York: John Wiley & Sons, 1976.

[146] B. Ata and J. A. Van Mieghem, "The Value of Partial Resource Pooling: Should a Service Network Be Integrated or Product-Focused?," *Manage. Sci.*, vol. 55, no. 1, pp. 115–131, Jan. 2009.

[147] A. Mandelbaum and M. I. Reiman, "On Pooling in Queueing Networks," *Manage. Sci.*, vol. 44, no. 7, pp. 971–981, Jul. 1998.

[148] N. Gans, G. Koole, and A. Mandelbaum, "Telephone Call Centers: Tutorial, Review, and Research Prospects," *Manuf. Serv. Oper. Manag.*, vol. 5, no. 2, pp. 79–141, Apr. 2003.

[149] P. B. Patel and D. R. Vinson, "Team Assignment System: Expediting Emergency Department Care," *Ann. Emerg. Med.*, vol. 46, no. 6, pp. 499–506, Dec. 2005.

[150] H. Song, A. L. Tucker, and K. L. Murrell, "The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay," *Manage. Sci.*, vol. 61, no. 12, pp. 3032–3053, Dec. 2015.

[151] J. M. Hirshon, T. D. Kirsch, W. K. Mysko, and G. D. Kelen, "Effect of rotational patient

assignment on emergency department length of stay," *J. Emerg. Med.*, vol. 14, no. 6, pp. 763–768, Nov. 1996.

[152] M. Shunko, J. Niederhoff, and Y. Rosokha, "Humans Are Not Machines: The Behavioral Impact of Queueing Design on Service Time," *Manage. Sci.*, vol. 64, no. 1, pp. 453–473, Jan. 2018.

[153] J. Wang and Y. P. Zhou, "Impact of queue configuration on service time: Evidence from a supermarket," *Manage. Sci.*, vol. 64, no. 7, pp. 3055–3075, Jul. 2018.

[154] G. P. Cachon and F. Zhang, "Obtaining Fast Service in a Queueing System via Performance-Based Allocation of Demand," *Manage. Sci.*, vol. 53, no. 3, pp. 408–420, Mar. 2007.

[155] S. M. Gilbert and Z. K. Weng, "Incentive Effects Favor Nonconsolidating Queues in a Service System: The Principal–Agent Perspective," *Manage. Sci.*, vol. 44, no. 12-part-1, pp. 1662–1669, Dec. 1998.

[156] W. J. Hopp, S. M. R. Iravani, and F. Liu, "Managing white-collar work: An operations-oriented survey," *Production and Operations Management*, vol. 18, no. 1. Wiley-Blackwell, pp. 1–32, 01-Jan-2009.

[157] W. J. Hopp, S. M. R. Iravani, and G. Y. Yuen, "Operations Systems with Discretionary Task Completion," *Manage. Sci.*, vol. 53, no. 1, pp. 61–77, Jan. 2007.

[158] O. Jouini, Y. Dallery, and R. Nait-Abdallah, "Analysis of the Impact of Team-Based Organizations in Call Center Management," *Manage. Sci.*, vol. 54, no. 2, pp. 400–414, Feb. 2008.

[159] T. F. Tan and S. Netessine, "When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity," *Manage. Sci.*, vol. 60, no. 6, pp. 1574–1593, Jun. 2014.

[160] H. T. Do, M. Shunko, M. T. Lucas, and D. C. Novak, "Impact of Behavioral Factors on Performance of Multi-Server Queueing Systems," *Prod. Oper. Manag.*, vol. 27, no. 8, pp. 1553–1573, Aug. 2018.

[161] M. Armony, G. Roels, and H. Song, "Pooling Queues with Discretionary Service Capacity," *SSRN Electron. J.*, no. May, 2017.

[162] S. Doroudi, R. Gopalakrishnan, and A. Wierman, "Dispatching to incentivize fast service in multi-server queues," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 39, no. 3, p. 43, Dec. 2011.

[163] D. J. France *et al.*, "Emergency physicians' behaviors and workload in the presence of an electronic whiteboard," *Int. J. Med. Inform.*, vol. 74, no. 10, pp. 827–837, Oct. 2005.

[164] M. Delasay, A. Ingolfsson, B. Kolfal, and K. Schultz, "Load effect on service times," *Eur. J. Oper. Res.*, vol. 279, no. 3, pp. 673–686, Dec. 2019.

[165] L. C. Edie, "Traffic Delays at Toll Booths," *J. Oper. Res. Soc. Am.*, vol. 2, no. 2, pp. 107–138, May 1954.

[166] L. G. Debo, L. B. Toktay, and L. N. Van Wassenhove, "Queuing for Expert Services," *Manage. Sci.*, vol. 54, no. 8, pp. 1497–1512, Aug. 2008.

[167] S. Hasija, E. Pinker, and R. A. Shumsky, "OM Practice —Work Expands to Fill the Time Available: Capacity Estimation and Staffing Under Parkinson's Law," *Manuf. Serv. Oper. Manag.*, vol. 12, no. 1, pp. 1–18, Jan. 2010.

[168] K. C. Diwas Singh, "Does multitasking improve performance? evidence from the emergency department," *Manuf. Serv. Oper. Manag.*, vol. 16, no. 2, pp. 168–183, May 2014.

[169] K. C. Diwas Singh and C. Terwiesch, "Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations," *Manage. Sci.*, vol. 55, no. 9, pp. 1486–1498, Sep. 2009.

[170] J. A. Berry Jaeker and A. L. Tucker, "Past the Point of Speeding Up: The Negative Effects of Workload Saturation on Efficiency and Patient Severity," *Manage. Sci.*, vol. 63, no. 4, pp. 1042–1062, Apr. 2017.

[171] L. Kuntz, R. Mennicken, and S. Scholtes, "Stress on the ward: Evidence of safety tipping points in hospitals," *Manage. Sci.*, vol. 61, no. 4, pp. 754–771, Apr. 2015.

[172] L. V. Green, S. Savin, and B. Wang, “Managing Patient Service in a Diagnostic Medical Facility,” *Oper. Res.*, vol. 54, no. 1, pp. 11–25, Feb. 2006.

[173] L. R. Schwartz and D. T. Overton, “Emergency department complaints: A one-year analysis,” *Ann. Emerg. Med.*, vol. 16, no. 8, pp. 857–861, Aug. 1987.

[174] E. L. Hahne, “Round-robin scheduling for max-min fairness in data networks,” *IEEE J. Sel. Areas Commun.*, vol. 9, no. 7, pp. 1024–1039, 1991.

[175] R. Oliva and J. D. Sterman, “Cutting corners and working overtime: Quality erosion in the service industry,” *Manage. Sci.*, vol. 47, no. 7, pp. 894–914, 2001.

[176] M. Armony and A. R. Ward, “Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems,” *Oper. Res.*, vol. 58, no. 3, pp. 624–637, Jun. 2010.

[177] N. Gans, G. Koole, A. Mandelbaum, N. Gans, G. Koole, and A. Mandelbaum, “Commissioned Paper Telephone Call Centers : Tutorial , Review , and Research Prospects,” no. September 2019, 2003.

[178] D. A. Stanford, P. Taylor, and I. Ziedins, “Waiting time distributions in the accumulating priority queue,” *Queueing Syst.*, vol. 77, no. 3, pp. 297–330, Jul. 2014.

[179] L. Mayhew and D. Smith, “Using queuing theory to analyse the Government’s 4-h completion time target in Accident and Emergency departments,” *Health Care Manag. Sci.*, vol. 11, no. 1, pp. 11–21, Mar. 2008.

[180] T. H. Taylor *et al.*, “A study of anaesthetic emergency work. Paper 1: The method of study and introduction of queuing theory,” *Br. J. Anaesth.*, vol. 41, no. 1, pp. 70–75, 1969.

[181] R. K. Haussmann, “Waiting time as an index of quality of nursing care.,” *Health Serv. Res.*, vol. 5, no. 2, pp. 92–105, 1970.

[182] K. Siddharthan, W. J. Jones, and J. A. Johnson, “A priority queuing model to reduce waiting times in emergency care,” *Int. J. Health Care Qual. Assur.*, vol. 9, no. 5, pp. 10–16, Sep. 1996.

[183] M. Laskowski, R. D. McLeod, M. R. Friesen, B. W. Podaima, and A. S. Alfa, "Models of Emergency Departments for Reducing Patient Waiting Times," *PLoS One*, vol. 4, no. 7, p. e6127, Jul. 2009.

[184] G. S. Mokaddis, I. A. Ismail, S. A. Metwally, and K. M. Metry, "Response times for health care system," *J. Appl. Math. Bioinforma.*, vol. 1, no. 2, pp. 131–146, 2011.

[185] D. G McQuarrie, "Hospitalization utilization levels. The application of queuing. Theory to a controversial medical economic problem," *Minn. Med.*, vol. 66, pp. 679–686, 1983.

[186] S. S. Panwalkar and W. Iskander, "A Survey of Scheduling Rules," *Oper. Res.*, vol. 25, no. 1, pp. 45–61, Feb. 2008.

[187] M. Armony, S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, and G. B. Yom-Tov, "On patient flow in hospitals: A data-based queueing-science perspective," *Stoch. Syst.*, vol. 5, no. 1, pp. 146–194, 2015.

[188] R. Hall, D. Belson, P. Murali, and M. Dessouky, "Modeling patient flows through the health care system," in *International Series in Operations Research and Management Science*, vol. 206, Hall R.W., Ed. Springer, Boston, MA, 2013, pp. 3–42.

[189] S. Zeltyn *et al.*, "Simulation-based models of emergency departments: Operational, Tactical and Strategic Staffing," *ACM Trans. Model. Comput. Simul.*, vol. 21, no. 4, pp. 1–25, Aug. 2011.

[190] I. for H. I. IHI, "Patient First: Efficient Patient Flow Management Impact on the ED," 2011. [Online]. Available: http://www.ihi.org/resources/Pages/ImprovementStories/PatientFirstEfficientPatientFlowManagementED.aspx. [Accessed: 01-Feb-2019].

[191] J. Huang, B. Carmeli, and A. Mandelbaum, "Control of Patient Flow in Emergency Departments, or Multiclass Queues with Deadlines and Feedback," *Oper. Res.*, vol. 63, no. 4, pp. 892–908, Aug. 2015.

[192] W. E. Smith, "Various optimizers for single-stage production," *Nav. Res. Logist. Q.*, vol. 3, no. 1–2, pp. 59–66, Mar. 1956.

[193] J. A. van Mieghem, “Dynamic Scheduling with Convex Delay Costs: The Generalized $c|mu$ Rule,” *Ann. Appl. Probab.*, vol. 5, no. 3, pp. 809–833, Aug. 2007.

[194] A. Mandelbaum and A. L. Stolyar, “Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized cμ-Rule,” *Oper. Res.*, vol. 52, no. 6, pp. 836–855, Dec. 2004.

[195] I. Gurvich and W. Whitt, “Scheduling Flexible Servers with Convex Delay Costs in Many-Server Service Systems,” *Manuf. Serv. Oper. Manag.*, vol. 11, no. 2, pp. 237–253, Apr. 2008.

[196] L. Kleinrock, “A delay dependent queue discipline,” *Nav. Res. Logist. Q.*, vol. 11, no. 3–4, pp. 329–341, Sep. 1964.

[197] A. Bin Sharif, D. A. Stanford, P. Taylor, and I. Ziedins, “A multi-class multi-server accumulating priority queue with application to health care,” *Oper. Res. Heal. Care*, vol. 3, no. 2, pp. 73–79, Jun. 2014.

[198] Y. B. Ferrand, M. J. Magazine, U. S. Rao, and T. F. Glass, “Managing responsiveness in the emergency department: Comparing dynamic priority queue with fast track,” *J. Oper. Manag.*, vol. 58–59, no. 1, pp. 15–26, Mar. 2018.

[199] G. Zayas-Caban, J. Xie, L. V. Green, and M. E. Lewis, “Policies for physician allocation to triage and treatment in emergency departments,” *IISE Trans. Healthc. Syst. Eng.*, pp. 1–15, Jun. 2019.

[200] P. Lindsay, “The Development of Indicators to Measure the Quality of Clinical Care in Emergency Departments Following a Modified-Delphi Approach,” *Acad. Emerg. Med.*, vol. 9, no. 11, pp. 1131–1139, Nov. 2002.

[201] L. I. Horwitz, J. Green, and E. H. Bradley, “US Emergency Department Performance on Wait Time and Length of Visit,” *Ann. Emerg. Med.*, vol. 55, no. 2, pp. 133–141, Feb. 2010.

[202] E. B. Kulstad, R. Sikka, R. T. Sweis, K. M. Kelley, and K. H. Rzechula, “ED overcrowding is associated with an increased frequency of medication errors,” *Am. J.*

Emerg. Med., vol. 28, no. 3, pp. 304–309, Mar. 2010.

[203] J. S. Olshaker and N. K. Rathlev, "Emergency Department overcrowding and ambulance diversion: The impact and potential solutions of extended boarding of admitted patients in the Emergency Department," *J. Emerg. Med.*, vol. 30, no. 3, pp. 351–356, Apr. 2006.

[204] S. J. Weiss, A. A. Ernst, M. R. Sills, B. J. Quinn, A. Johnson, and T. G. Nick, "Development of a Novel Measure of Overcrowding in a Pediatric Emergency Department," *Pediatr. Emerg. Care*, vol. 23, no. 9, pp. 641–645, Sep. 2007.

[205] F. Mallor, C. Azcárate, and J. Barado, "Control problems and management policies in health systems: application to intensive care units," *Flex. Serv. Manuf. J.*, vol. 28, no. 1–2, pp. 62–89, Jun. 2016.

[206] C. M. Macal *et al.*, "Modeling the spread of community-associated MRSA," in *Proceedings Title: Proceedings of the 2012 Winter Simulation Conference (WSC)*, 2012, no. 2005, pp. 1–12.

[207] F. Mallor and C. Azcárate, "Combining optimization with simulation to obtain credible models for intensive care units," *Ann. Oper. Res.*, vol. 221, no. 1, pp. 255–271, Oct. 2014.

[208] M. Laguna and R. Martí, "The OptQuest Callable Library," in *Optimization Software Class Libraries*, vol. 18, no. 42743, Boston: Kluwer Academic Publishers, 2005, pp. 193–218.

[209] N. Li and D. A. Stanford, "Multi-server accumulating priority queues with heterogeneous servers," *Eur. J. Oper. Res.*, vol. 252, no. 3, pp. 866–878, Aug. 2016.

[210] P. Knauth, "Designing better shift systems," in *Applied Ergonomics*, 1996, vol. 27, no. 1, pp. 39–44.

[211] J. Van den Bergh, J. Beliën, P. De Bruecker, E. Demeulemeester, and L. De Boeck, "Personnel scheduling: A literature review," *Eur. J. Oper. Res.*, vol. 226, no. 3, pp. 367–385, May 2013.

[212] P. De Bruecker, J. Van den Bergh, J. Beliën, and E. Demeulemeester, "Workforce

planning incorporating skills: State of the art," *Eur. J. Oper. Res.*, vol. 243, no. 1, pp. 1–16, May 2015.

[213] E. K. Burke, P. De Causmaecker, G. Vanden Berghe, and H. Van Landeghem, "The state of the art of nurse rostering," *J. Sched.*, vol. 7, no. 6, pp. 441–449, Nov. 2004.

[214] B. Cheang, H. Li, A. Lim, and B. Rodrigues, "Nurse rostering problems—a bibliographic survey," *Eur. J. Oper. Res.*, vol. 151, no. 3, pp. 447–460, Dec. 2003.

[215] M. Erhard, J. Schoenfelder, A. Fügener, and J. O. Brunner, "State of the art in physician scheduling," *Eur. J. Oper. Res.*, vol. 265, no. 1, pp. 1–18, Feb. 2018.

[216] J. O. Brunner, J. F. Bard, and R. Kolisch, "Flexible shift scheduling of physicians," *Health Care Manag. Sci.*, vol. 12, no. 3, pp. 285–305, Sep. 2009.

[217] J. O. Brunner, J. F. Bard, and R. Kolisch, "Midterm scheduling of physicians with flexible shifts using branch and price," *IIE Trans.*, vol. 43, no. 2, pp. 84–109, Nov. 2010.

[218] Ministerio de Empleo y Seguridad Social, *Boletín Oficial Del Estado, BOE-A-2015-11430*. 2015, pp. 100224–100308.

[219] S. Topaloglu, "A shift scheduling model for employees with different seniority levels and an application in healthcare," *Eur. J. Oper. Res.*, vol. 198, no. 3, pp. 943–957, Nov. 2009.

[220] H. Beaulieu, J. A. Ferland, B. Gendron, and P. Michelon, "A mathematical programming approach for scheduling physicians in the emergency room," *Health Care Manag. Sci.*, vol. 3, no. 3, pp. 193–200, 2000.

[221] R. M. Karp, "Reducibility among Combinatorial Problems," in *Complexity of Computer Computations*, Boston, MA: Springer US, 1972, pp. 85–103.

[222] A. T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier, "Staff scheduling and rostering: A review of applications, methods and models," *Eur. J. Oper. Res.*, vol. 153, no. 1, pp. 3–27, 2004.

[223] J. Puente, A. Gómez, I. Fernández, and P. Priore, "Medical doctor rostering problem in

a hospital emergency department by means of genetic algorithms," *Comput. Ind. Eng.*, vol. 56, no. 4, pp. 1232–1242, May 2009.

[224] H. Vermuyten, J. Namorado Rosa, I. Marques, J. Beliën, and A. Barbosa-Póvoa, "Integrated staff scheduling at a medical emergency service: An optimisation approach," *Expert Syst. Appl.*, vol. 112, pp. 62–76, Dec. 2018.

[225] R. C. Carrasco, "Long-term staff scheduling with regular temporal distribution," *Comput. Methods Programs Biomed.*, vol. 100, no. 2, pp. 191–199, Nov. 2010.

[226] T. A. Feo and M. G. . Resende, "A probabilistic heuristic for a computationally difficult set covering problem," *Oper. Res. Lett.*, vol. 8, no. 2, pp. 67–71, Apr. 1989.

[227] T. A. Feo and M. G. C. Resende, "Greedy Randomized Adaptive Search Procedures," *J. Glob. Optim.*, vol. 6, no. 2, pp. 109–133, Mar. 1995.

[228] J. L. Bresina, "Heuristic-biased stochastic sampling," in *AAAI/IAAI, Vol. 1*, 1996, pp. 271–278.

[229] P. Festa and M. G. C. Resende, "Grasp: An Annotated Bibliography," Springer, Boston, MA, 2002, pp. 325–367.

[230] N. Mladenović and P. Hansen, "Variable neighborhood search," *Comput. Oper. Res.*, vol. 24, no. 11, pp. 1097–1100, Nov. 1997.

[231] P. Hansen and N. Mladenović, "Variable neighborhood search: Principles and applications," *Eur. J. Oper. Res.*, vol. 130, no. 3, pp. 449–467, May 2001.

[232] A. Duarte, J. Sánchez-Oro, N. Mladenović, and R. Todosijević, "Variable Neighborhood Descent," in *Handbook of Heuristics*, Cham: Springer International Publishing, 2018, pp. 341–367.

[233] R. K. Ahuja, Ö. Ergun, J. B. Orlin, and A. P. Punnen, "A survey of very large-scale neighborhood search techniques," *Discret. Appl. Math.*, vol. 123, no. 1–3, pp. 75–102, Nov. 2002.

[234] A. P. Punnen, "The traveling salesman problem: new polynomial approximation

algorithms and domination analysis," *J. Inf. Optim. Sci.*, vol. 22, no. 1, pp. 191–206, Jan. 2001.

[235] M. Dror and L. Levy, "A vehicle routing improvement algorithm comparison of a 'greedy' and a matching implementation for inventory routing," *Comput. Oper. Res.*, vol. 13, no. 1, pp. 33–45, Jan. 1986.

[236] D. M. Warner, "Nurse staffing, scheduling, and reallocation in the hospital.," *Hosp. Health Serv. Adm.*, vol. 21, no. 3, pp. 77–90, 1976.

[237] A. De Kreuk, E. Winands, and J. Vissers, "Master scheduling of medical specialists," in *Health Operations Management: Patient Flow Logistics in Health Care*, vol. 13, Routledge, 2005, pp. 184–201.

[238] A. Gunawan and H. C. Lau, "Master physician scheduling problem," *J. Oper. Res. Soc.*, vol. 64, no. 3, pp. 410–425, Mar. 2013.

[239] J. F. Bard, Z. Shu, and L. Leykum, "Monthly clinic assignments for internal medicine housestaff," *IIE Trans. Healthc. Syst. Eng.*, vol. 3, no. 4, pp. 207–239, Oct. 2013.

[240] E. W. Hans, M. van Houdenhoven, and P. J. H. Hulshof, "A Framework for Healthcare Planning and Control," in *Handbook of Healthcare System Scheduling*, vol. 168, no. January 2012, R. Hall, Ed. Boston, MA: Springer Boston, MA, 2012, pp. 303–320.

[241] G. Thopsn, "Labor Scheduling, Part 3: Developing a Workforce Schedule," *Cornell Hotel Restaur. Adm. Q.*, vol. 40, no. 1, pp. 86–96, Jun. 1999.

[242] G. Koole and A. Pot, "An overview of routing and staffing algorithms in multi-skill customer contact centers," *Submitt. Publ.*, vol. 39, no. 6, pp. 1–42, Dec. 2006.

[243] G. Thompson, "Labor Scheduling, Part 1: Forecasting Demand," *Cornell Hotel Restaur. Adm. Q.*, vol. 39, no. 5, pp. 22–31, Oct. 1998.

[244] G. M. Thompson, "Labor scheduling, part 2: Knowing how many on-duty employees to schedule," *Cornell Hotel Restaur. Adm. Q.*, vol. 39, no. 6, pp. 26–37, 1998.

[245] E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "AN INTEGRATED WORK SHIFT

SCHEDULING SYSTEM," *Decis. Sci.*, vol. 7, no. 4, pp. 620–630, Oct. 1976.

[246] G. M. Thompson, "Labor scheduling using NPV estimates of the marginal benefit of additional labor capacity," *J. Oper. Manag.*, vol. 13, no. 1, pp. 67–86, Jul. 1995.

[247] M. Defraeye and I. Van Nieuwenhuyse, "Staffing and scheduling under nonstationary demand for service: A literature review," *Omega (United Kingdom)*, vol. 58. Elsevier Ltd, pp. 4–25, 01-Jan-2016.

[248] T. A. Feo, M. G. C. Resende, and S. H. Smith, "A Greedy Randomized Adaptive Search Procedure for Maximum Independent Set," *Oper. Res.*, vol. 42, no. 5, pp. 860–878, Oct. 1994.

[249] K. "KT" Thulasiraman *et al.*, *Handbook of Graph Theory, Combinatorial Optimization, and Algorithms*. Chapman and Hall/CRC, 2015.

[250] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows : Theory, Algorithms, and Applications*. Upper Saddle River, NJ, USA: Prentice Hall, 1993.

Appendix A Instructions sheet for the completion of the stress questionnaire by the experts

Description of the stress questionnaire and instruction sheet for completing it provided to expert raters in the training session.

upna Gobierno de Navarra

Research Group DECYL

ASSESSMENT OF STRESS DUE TO INSTANTANEOUS WORKLOAD

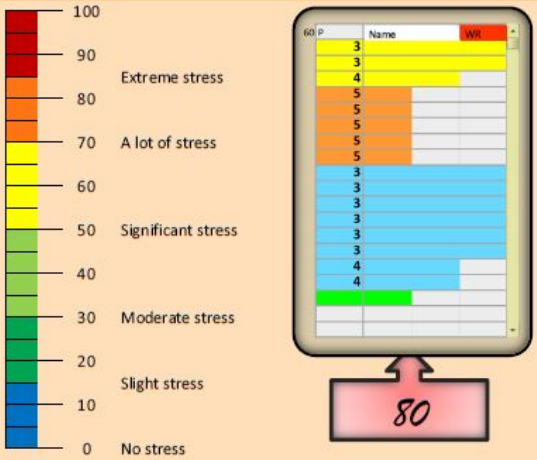
The following survey evaluates the **stress** perceived by physicians due to their assigned **workload at a given time**, which is called the **instantaneous workload**. An instantaneous workload is considered the **set of pending patients a doctor has been assigned simultaneously** – distinguishing their priority, as well as the **medical attention phase** of the patients– and his/her training responsibility.

To carry out this assessment, physicians will be presented several scenarios of different pending workloads associated to a physician in the same way as the patients' portfolio of the emergency department computer. This board lets physicians be aware of all the pending patients they have been assigned to them as patients are triaged and then immediately assigned to a specific physician as they arrive to the Emergency Department. The following colour code is used:

Priority	Name	WoutR / WR
3, 4, 5	Patients waiting to be seen for the first time.	
3, 4, 5	Patients waiting to be seen for the first time, they have exceeded the waiting time limit.	
3, 4, 5	Patients who have already been seen by the doctor, who have requested medical tests. These patients are still inside the ED system waiting for results and being discharge.	
3, 4, 5	Patients who have receive the medical discharge, but remain in the emergency department waiting to be transferred to the hospital, home, etc. In case of these patients to get worse, the assigned physician is still responsible for your care	

Patients who have been discharged and have left the system are not considered as pending workload.

What follows is a set of scenarios which show several patients a doctor has been assigned at a specific moment of his/her work-shift, and whether ***the shift*** he/she is working does or does not ***involve the supervision of a first-year resident*** (indicated in the red colour at the top right corner of the computer screen represented above. The experts should assess the perceived stress associated to each of the scenarios. This will involve that experts give a score between 0 and 100, helped by the qualitative scale described below.



Scenario:
The physician has been assigned three patients whose waiting time limit has been exceeded (two of priority 3 and one of priority 4). In addition, there are also five patients with minimum severity index (5) to be seen, and eight patients (six with priority 3) to receive a second consultation and be discharged. Moreover, the physician has a patient who has got his/her ED process finished but who is waiting to be transferred to another hospital or home. Finally, he has to supervise a resident.

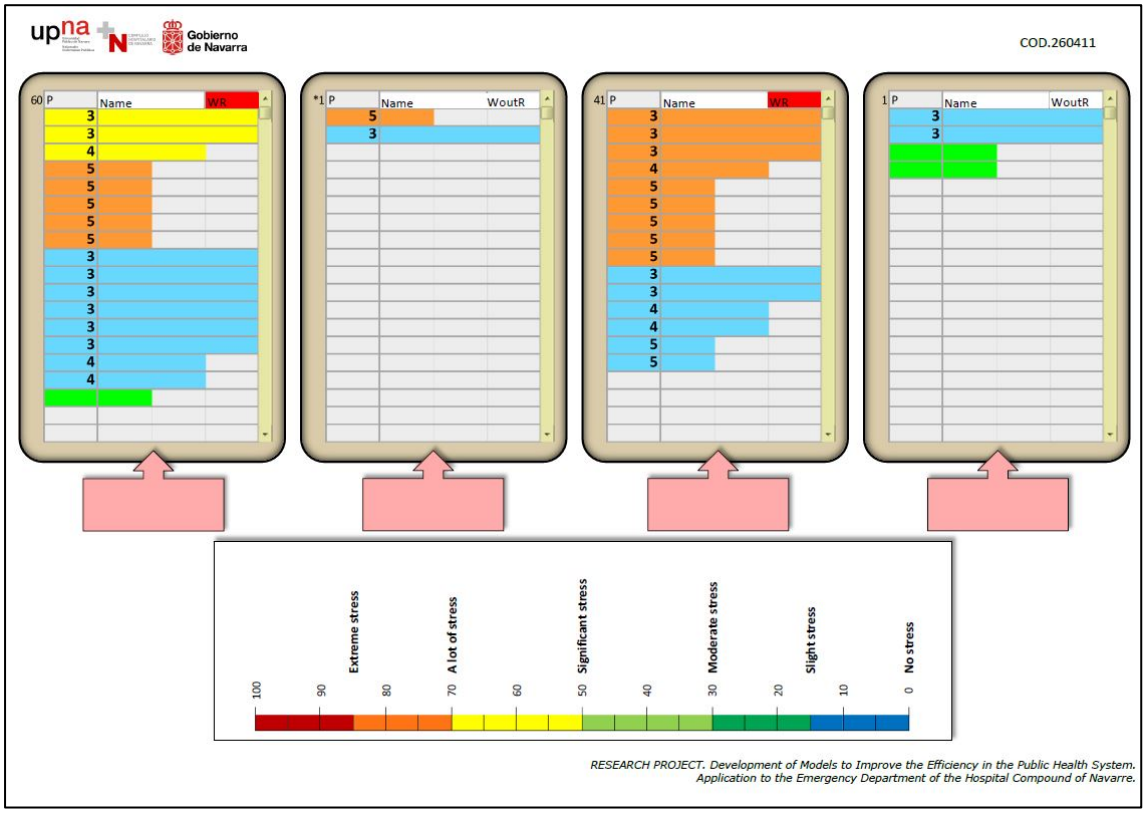
Physician Stress Assessment:
If the physician considers that the stress produced by the workload assigned in his/her patients' portfolio is very high, he/she could assess the stress situation with a score of 80.

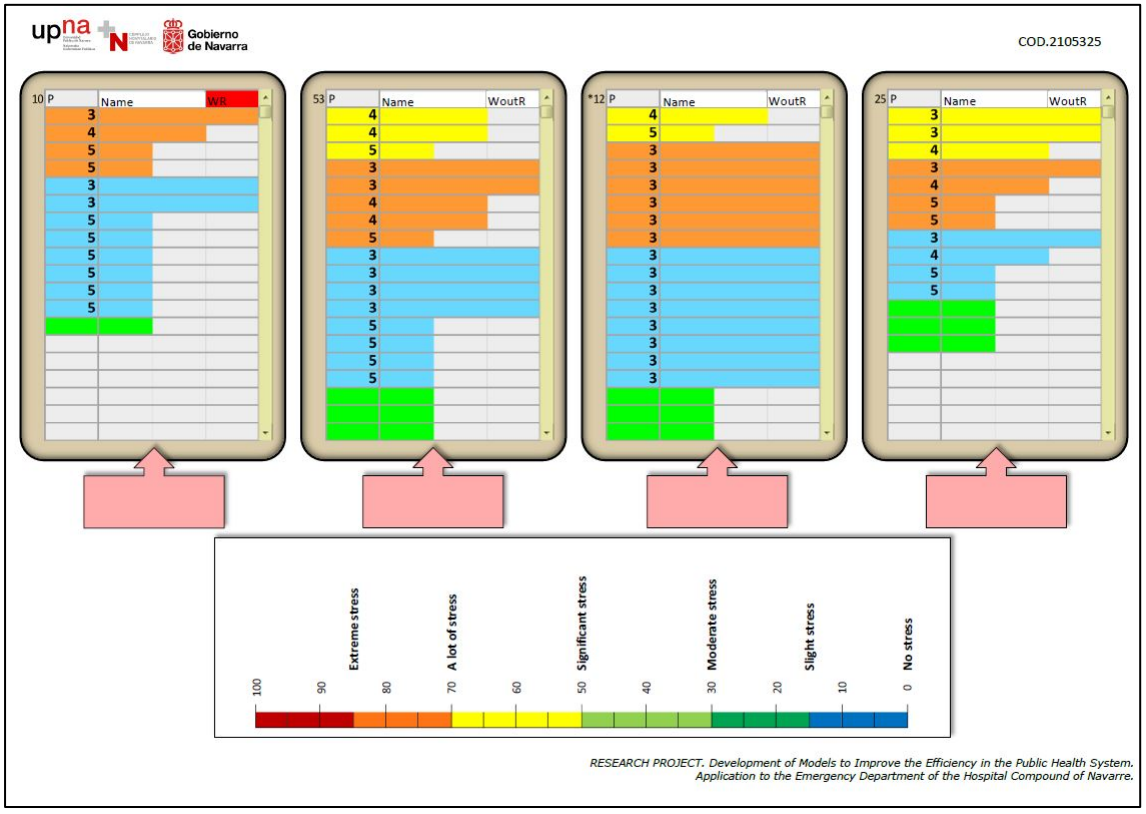
***** Note: There are no correct or incorrect ratings. You are asked to provide a subjective assessment of the stress due to workload assigned to the physician at a given point in the work shift.***

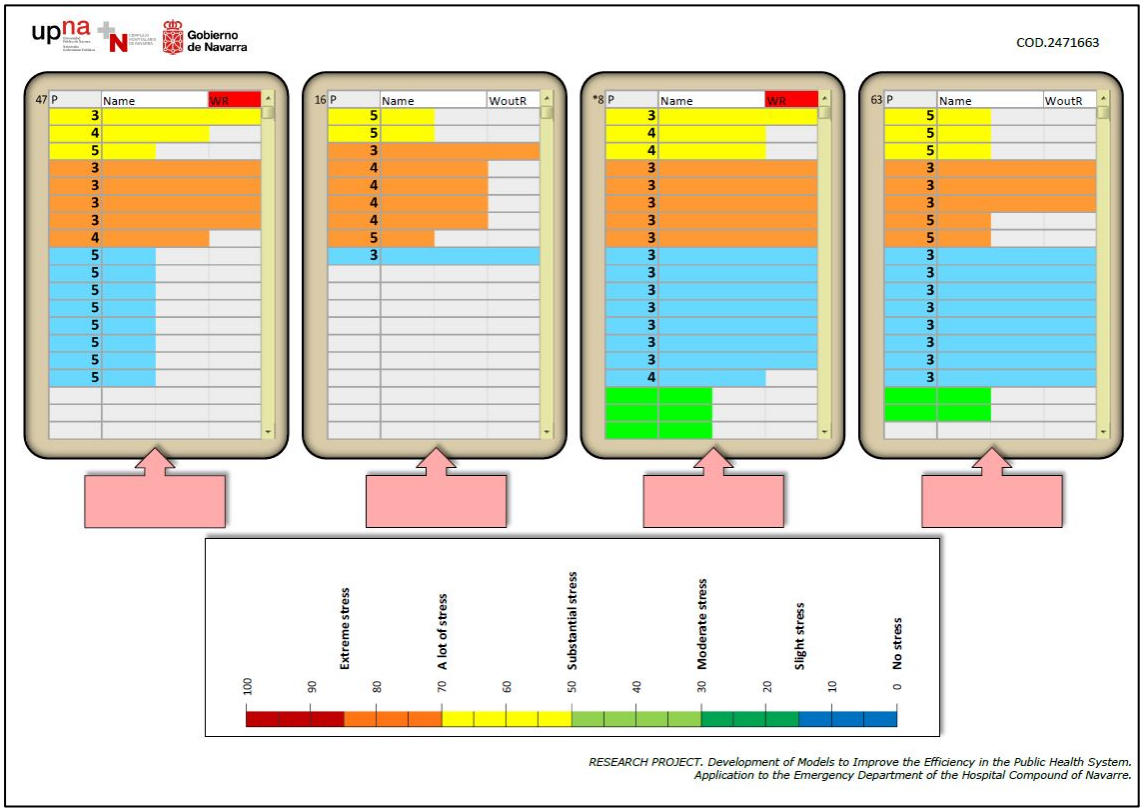
RESEARCH PROJECT. Development of Models to Improve the Efficiency in the Public Health System. Application to the Emergency Department of the Hospital Compound of Navarre.

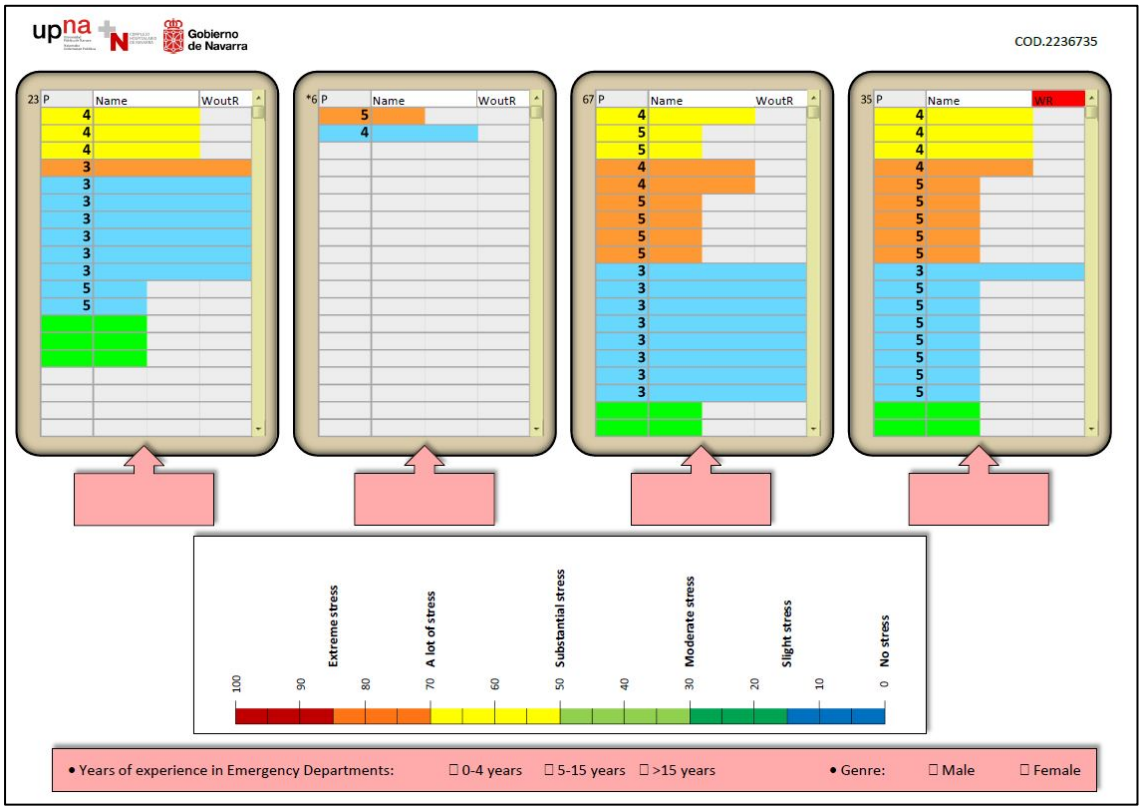
Appendix B Stress questionnaire example

In this appendix, we show one of the six stress questionnaires designed. They only differ in the set of scenarios provided in each of the four cards to be assessed in terms of stress.









Appendix C Consistency with the group index, CGI, for inter-respondent consistency analysis

A group of physicians is defined by those that completed the same questionnaire, and consequently, have assessed the stress of the same scenarios.

A group of n physicians is denoted as $A = \{D_1, \dots, D_i, \dots, D_n\}$, the set of m scenarios forming the questionnaire answered by the group A as $\Omega_A = \{S_1, \dots, S_j, \dots, S_m\}$ $(\Omega_A \subseteq \Omega)$ and the stress score of a physician i for scenario j $Y_i(S_j) = Y_{ij}$. We construct the matrix, M_i, which indicates above the main diagonal the stress comparisons between scenarios made by physician i.

$$M_i = [m_i(j,k)] \;\; such\ that \;\; \forall j \geq k \; m_i(j,k) = 0 \;\; and \;\; \forall j < k \;\; m_i(j,k)$$
$$= \begin{cases} 1 \; if \; Y_{ij} < Y_{ik} \\ 0 \; if \; Y_{ij} = Y_{ik} \\ -1 \; if \; Y_{ij} > Y_{ik} \end{cases}$$

Another matrix Q_A is defined to reflect the consensus of group A of physicians about their stress comparisons between scenarios.

$$Q_A = [q_A(j,k)] \;\; such\ that \;\; q_A(j,k) = \begin{cases} 1 \; if \; (\sum_{i \in A} m_i(j,k)) > 0 \\ 0 \; if \; (\sum_{i \in A} m_i(j,k)) = 0 \\ -1 \; if \; (\sum_{i \in A} m_i(j,k)) < 0 \end{cases} \quad \forall \; i \neq h$$

The agreements of a physician h with the rest of physicians in his/her group A are stored in a matrix $G_{A(h)}$, defined from the matrices M_i and $Q_{A_{-h}}$, where A_{-h} denotes the set A minus the physician h ($A_{-h} = A - \{h\}$):

$$G_{A(h)} = [g_{A(h)}(j,k)] \;\; such\ that \;\; \forall j \geq k \;\; g_{A(h)}(j,k) = 0 \;\; and$$

$$\forall j < k \quad g_{A(h)}(j,k) = \begin{cases} 1\ if\ \ m_i(j,k) = q_{A_{-h}}(j,k), & concordance \\ 0\ if\ \ q_{A_{-h}}(j,k) = 0\ and\ \ m_i(j,k) \neq 0, & indecisiveness \\ -1\ if\ \ q_{A_{-h}}(j,k) = -\,m_i(j,k), & discordance \end{cases}$$

The values $g_{A(h)}(j,k)$ reflect three situations between a physician h and the rest of the group:

Concordance: physician and the rest of the group assigned the same order to a pair of scenarios S_j, S_k from most stressful to less stressful.

Indecisiveness: one half and the other half of the group members' assigned the opposite order to a pair of scenarios S_j, S_k from most stressful to less stressful. Physician h would break the tie among the rest of the group physicians.

Discordance: physician and the rest of the group assigned the opposite order to a pair of scenarios S_j, S_k from most stressful to less stressful.

Finally, the consistency with the group index, CGI, taking into consideration the number of concordances ($C = \sum_{j,k} 1_{\{g_{A(h)}(j,k)=1\}}$), discordances ($D = \sum_{j,k} 1_{\{g_{A(h)}(j,k)=-1\}}$), and indecisiveness ($I = \sum_{j,k} 1_{\{g_{A(h)}(j,k)=0\}}$) in the matrix $G_{A(h)}$ is defined as:

$$CGI_{A(h)} = \frac{(C-D)}{(C+D+I)}$$

An "inconsistent rater" – who should be excluded for the study – is a rater whose CGI is below a certain limit $L_l \in \mathbb{R}$. In this study, we consider $L_l = 0.25$.

Appendix D Instructions sheet for the completion of the workload completed questionnaire by the experts

Description of the workload completed questionnaire and instruction sheet for completing it provided to expert raters in the training session.

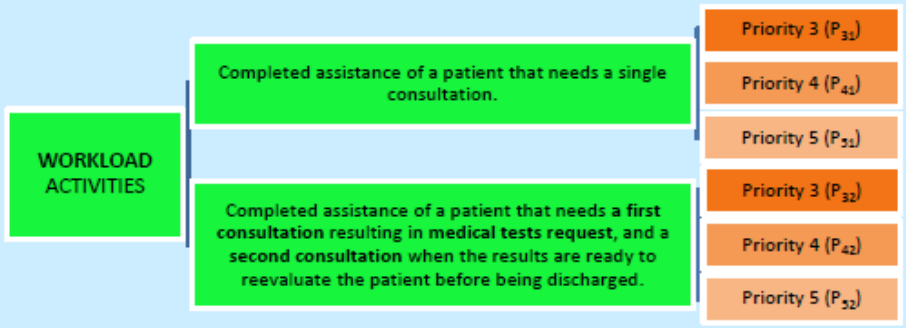
upna Gobierno de Navarra

Grupo de Investigación DECYL

ASSESSMENT OF WORKLOAD COMPLETED DURING A WORKSHIFT

The term *workload* associated to a specific physician intervention refers to "**servers capacity actually required to develop a specific task.**" *(O'Donnell and Eggemeier, 1986)*. This definition includes the time needed, physician intensity of the work and mental effort required to the task performance.

The **workload achieved by a physician during a work shift** is a function of the **number of patients assisted** of each of the following 6 groups.



The workload assessment is conducted through **the comparison of the attention required by two different types of patients.** Each of the 15 existing comparisons is represented by a row of the questionnaire table in which it is necessary to tick a box. The number associated to each column of the row represents how greater the workload associated to the complete assistance of a patient is compared to the other patient's, on a scale of 1 to 9.

The following example shows how to fill in each row in the table:

Example. When comparing the workload associated to the complete assistance of a priority 3 patient of type P32 to that associated to the complete assistance of a priority 5 patient of type P_{51}, if we consider that the first is extremely greater than the second, then we would tic the box closest to P_{32} patient associated to number 9.

	PATIENT TYPE	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	PATIENT TYPE
8	*P3, consultation + + medical tests + + consultation. (P_{32})*	X																	*P5, single consultation. (P_{51})*

Scale	Definition
1	The complete assistance of both types of patients require **equal** workload.
2	
3	The complete assistance of the patient type closest to the box implies **moderately greater** workload than the other's.
4	
5	The complete assistance of the patient type closest to the box implies **essentially greater** workload than the other's.
6	
7	The complete assistance of the patient type closest to the box implies **much greater** workload than the other's.
8	
9	The complete assistance of the patient type closest to the box implies **extremely greater** workload than the other's.

* 2, 4, 6 y 8 are intermediate values between the two adjacent judgements.

Appendix E Workload completed questionnaire example

upna Gobierno de Navarra

Grupo de Investigación DECYL

	PATIENT TYPE	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	PATIENT TYPE
1	P3, single consultation. (P_{31})																		P3, consultation + + medical tests + + consultation. (P_{32})
2	P3, single consultation. (P_{31})																		P4, single consultation. (P_{41})
3	P3, single consultation. (P_{31})																		P4, consultation + + medical tests + + consultation. (P_{42})
4	P3, single consultation. (P_{31})																		P5, single consultation. (P_{51})
5	P3, single consultation .(P_{31})																		P5, consultation + + medical tests + + consultation. (P_{52})
6	P3, consultation + + medical tests + + consultation. (P_{32})																		P4, single consultation. (P_{41})
7	P3, consultation + + medical tests + + consultation. (P_{32})																		P4, consultation + + medical tests + + consultation. (P_{42})
8	P3, consultation + + medical tests + + consultation. (P_{32})																		P5, single consultation. (P_{51})

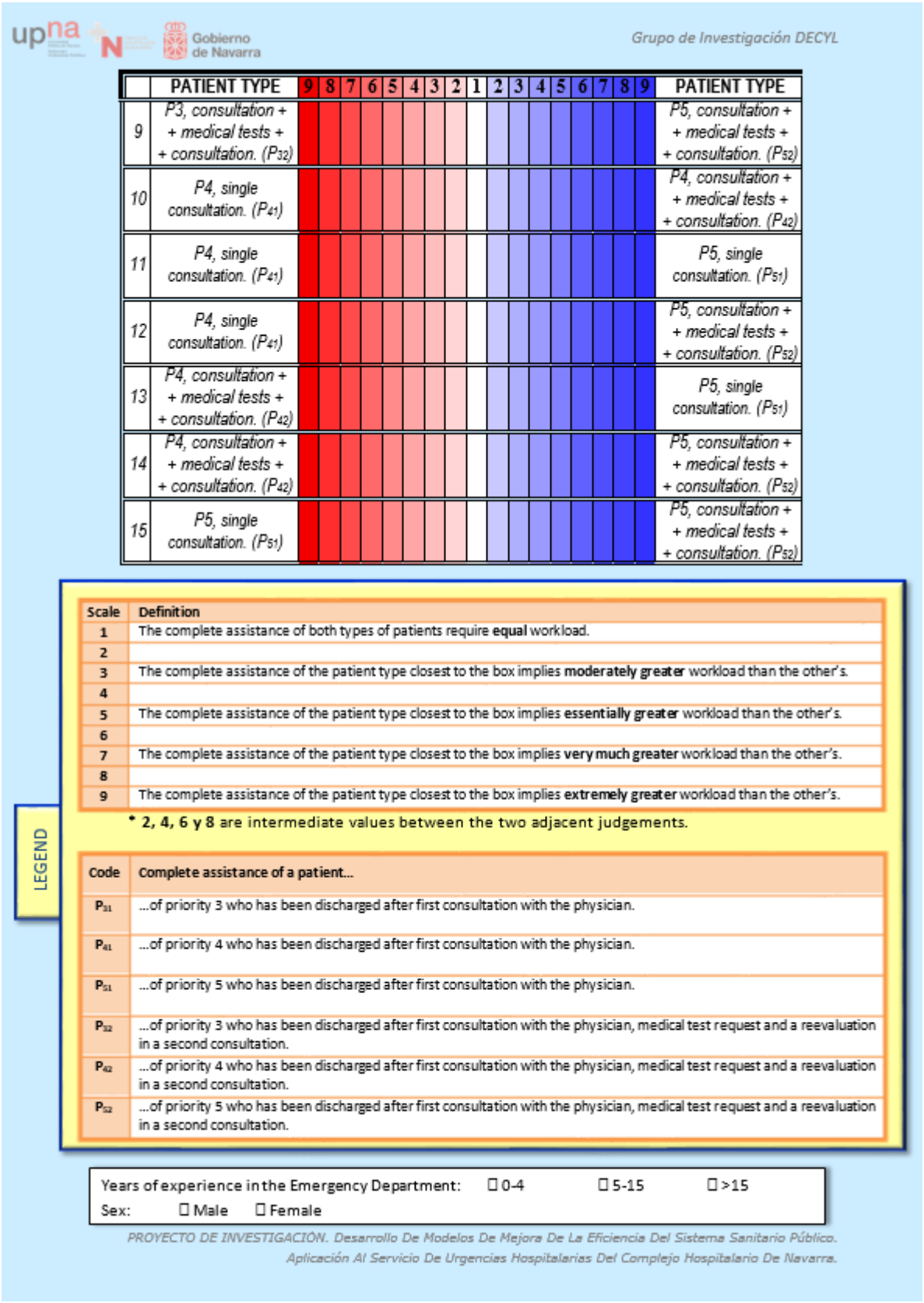
LEGEND

Scale	Definition
1	The complete assistance of both types of patients require **equal** workload.
2	
3	The complete assistance of the patient type closest to the box implies **moderately greater** workload than the other's.
4	
5	The complete assistance of the patient type closest to the box implies **essentially greater** workload than the other's.
6	
7	The complete assistance of the patient type closest to the box implies **much greater** workload than the other's.
8	
9	The complete assistance of the patient type closest to the box implies **extremely greater** workload than the other's.

* **2, 4, 6 y 8** are intermediate values between the two adjacent judgements.

Code	Complete assistance of a patient...
P_{31}	...of priority 3 who has been discharged after first consultation with the physician.
P_{41}	...of priority 4 who has been discharged after first consultation with the physician.
P_{51}	...of priority 5 who has been discharged after first consultation with the physician.
P_{32}	...of priority 3 who has been discharged after first consultation with the physician, medical test request and a reevaluation in a second consultation.
P_{42}	...of priority 4 who has been discharged after first consultation with the physician, medical test request and a reevaluation in a second consultation.
P_{52}	...of priority 5 who has been discharged after first consultation with the physician, medical test request and a reevaluation in a second consultation.

PROYECTO DE INVESTIGACIÓN. Desarrollo De Modelos De Mejora De La Eficiencia Del Sistema Sanitario Público. Aplicación Al Servicio De Urgencias Hospitalarias Del Complejo Hospitalario De Navarra.



Appendix F Notation and table of acronyms of Part I

Acronym	Definition
General terms	
ED	Emergency Department
KPI	Key performance indicator
CTAS	Canadian Triage Acuity Scale
OFV	Objective function value
SBO	Simulation based optimization
DES	Discrete event simulation model
Management policies	
APQ	Accumulative priority queue management policy
APQ-h	Accumulative priority queue with a finite horizon management policy (an extension of the normal APQ)
PP	Priority points
β_{1i}	Rate at which patients of class *i* who are waiting for the first consultation accumulate PP
β_{2i}	Rate at which patients of class *i* who are waiting for the second consultation accumulate PP
PR	Pure priority rule
FCFS	First come first served management policy
FIFO	First in first out management policy
PR-1C	1st Consultation pure priority rule
PR-2C	2nd Consultation pure priority rule
PR-AI	The acuity index pure priority rule
PR-HN	The rule which is used by the majority of the medical stuff in the HCN

Key performance indicators	
APT	Arrival to provider time (“door to doc”)
LoS	Length of stay
TWT	Total waiting time
Classes of patients	
1C	Patients waiting for the first consultation
2C	Patients waiting for the second consultation
HP	High-priority patients
MP	Medium-priority patients
LP	Low-priority patients
1C-HP	High-priority patients waiting for the first consultation
1C-MP	Medium-priority patients waiting for the first consultation
1C-LP	Low-priority patients waiting for the first consultation
2C-HP	High-priority patients waiting for the second consultation
2C-MP	Medium-priority patients waiting for the second consultation
2C-LP	Low-priority patients waiting for the second consultation
P3	Priority 3 patients
P4	Priority 4 patients
P5	Priority 5 patients
Scenario factors and levels	
F1	ED congestion level. It is the average occupation rate, ρ, $f_1 = \{90\%, 95\%\}$
F2	Arrival ($\lambda(t)$) seasonality, $f_2 = \{T0,\ Tu, Tp\}$
F3	Mix of patients, $f_3 = \{B0, B3, B4, B5\}$
T0	Constant arrival rate of patients λ(t)
Tu	Triangular pattern for the arrival rate λ(t), with a peak at 11:30 a.m. and a ratio of $(\lambda_{max} - \lambda_{min})/\lambda_{min} = 0.5$. It extends the triangular shape across the entire time range
Tp	Triangular pattern for the arrival rate λ(t), with a peak at 11:30 a.m. and a ratio of $(\lambda_{max} - \lambda_{min})/\lambda_{min} = 0.5$. It only applies the triangular shape in the time range [10:00, 13:00], with the arrival rate out of this range being constant
B0	Equilibrated. Balanced distribution among all types of patients (1/3 of P3, 1/3 of P4 and 1/3 of P5)

B3	Biased mix of patients towards priority 3 patients (50% of P3, 25% of P4 and 25% of P5)
B4	Biased mix of patients towards priority 4 patients (25% of P3, 50% of P4 and 25% of P5)
B5	Biased mix of patients towards priority 5 patients (25% of P3, 25% of P4 and 50% of P5)

Appendix G Assessment surveys for Chapter 4 pilot test

This appendix presents the survey used to assess the outcome of the pilot testing of the new allocation rule in the HCN previously evaluated by means of simulation (Chapter 4)

ENCUESTA DE VALORACIÓN

Programa **APAMET** (Asignación Paciente Médico en Triaje)

Este cuestionario anónimo persigue la valoración por parte del personal de enfermería de triaje de la aplicación APAMET (Asignación de Paciente a Médico en Triaje), cuya prueba piloto se ha desarrollado en el triaje del Servicio de Urgencias desde el 4 de junio hasta el 1 de julio de 2018.

- Número de días que ha trabajado con la aplicación en triaje durante el mes de junio:

❒ 1 ❒ 2 ❒ 3 ❒ 4 ❒ 5 ❒ 6 ❒ 7 ❒ 8 ❒ 9 ❒ 10 ❒ 11 ❒ 12 ❒ 13 ❒ 14 ❒ 15 ❒ >15

- Número de años de experiencia en urgencias: ❒ 1-2 ❒ 2-5 ❒ >5
- Sexo: ❒ Masculino ❒ Femenino
- Tipo de jornada en urgencias
  - ❒ Completa
  - ❒ 1/3
  - ❒ 1/2
  - ❒ Canguro (40%)
  - ❒ Otra: .....

A. Por favor, valore la conveniencia de introducir las siguientes mejoras al programa del 1 al 5 donde *1* es *"Nada conveniente"* y *5* es *"Muy conveniente"*.

	1	2	3	4	5
1. Integración con el Historial Clínico (evitaría la introducción manual de información del paciente)	❒	❒	❒	❒	❒
2. Desconexión parcial de consultas durante los turnos (por ausencia temporal de médicos)	❒	❒	❒	❒	❒

3. Indique, por favor, alguna otra modificación del programa que en su opinión mejoraría sus prestaciones:

B. Por favor, valore el grado de acuerdo o desacuerdo de las siguientes afirmaciones 1 al 5 donde *1* es *"Totalmente en desacuerdo"* y *5* es *"Totalmente de acuerdo"*.

Cuestiones sobre utilidad	1	2	3	4	5
1. La asignación de paciente a consulta resulta más sencilla con el programa que con la asignación manual.	❒	❒	❒	❒	❒
Sencillez de uso	1	2	3	4	5
2. El programa es sencillo y no requiere mucho tiempo para aprender a manejarlo.	❒	❒	❒	❒	❒
3. Las pantallas muestran claramente la información requerida que se debe introducir.	❒	❒	❒	❒	❒
4. ¿Ha tenido algún problema técnico durante su uso? ❒SI ❒ NO					
En caso afirmativo, la resolución de problemas por parte de la asistencia técnica durante el período de prueba ha sido rápida.	❒	❒	❒	❒	❒

C. Conteste a las siguientes preguntas (SÍ / NO)

Recomendación	SI	NO
5. ¿Recomendaría el desarrollo de una aplicación similar para el Circuito B?	❒	❒
6. ¿Recomendaría que el programa se implantase una vez introducidas las mejoras e integrado con el Historial Clínico?	❒	❒

Appendix H Additional numerical results of Chapter 5

Table 9.1 displays each scenario detailed results by disclosing the value for every KPI considered: the policy applied is in first column, the objective function value is in second column, the time target objectives values for P3, P4, and P5 priority patients are in columns third ($E(X_3)$), fourth ($E(X_4)$) and fifth ($E(X_5)$) respectively, the APT for P3, P4, and P5 are in sixth ($E(\tau_5)$), seventh ($E(\tau_4)$), and eighth ($E(\tau_5)$) columns and the TWT for P3, P4, and P5 patients who need two consultations are in ninth ($E(\tau_3 + \nu_3)$), tenth ($E(\tau_4 + \nu_4)$), and eleventh ($\mathrm{E}(\tau_5 + \nu_5)$) respectively. The last two columns are the description of each scenario and the improvement of the objective function value with respect to the best Pure Priority Rule.

Queue Discipline	Obj	$E(X_3)$	$E(X_4)$	$E(X_5)$	$E(\tau_3)$	$E(\tau_4)$	$E(\tau_5)$	$E(\tau_3 + \nu_3)$	$E(\tau_4 + \nu_4)$	$E(\tau_5 + \nu_5)$	Scenario	APQ-h improve-ment
PR-AI	106.15	0.00	<0.01	0.02	2.06	4.44	17.99	4.88	12.08	68.84		
PR-1C	145.03	0.00	<0.01	<0.01	2.10	3.24	6.52	18.01	33.27	68.69	F1: 90%;	0%
PR-2C	75.27	<0.01	<0.01	0.06	3.87	7.80	30.90	5.81	10.01	33.50	F2:T0;	
PR-HN	80.72	0.00	0.01	0.06	2.00	8.18	31.25	4.88	11.61	35.61	F3: B0	
APQ-h	75.28	<0.01	<0.01	0.06	3.47	7.79	30.99	5.63	9.90	34.23		
PR-AI	107.83	<0.01	0.01	0.05	2.34	7.71	26.39	6.59	19.91	79.42		
PR-1C	161.07	<0.01	<0.01	<0.01	2.38	4.16	7.39	22.88	46.64	78.84	F1: 90%;	
PR-2C	82.13	<0.01	0.03	0.10	4.71	11.85	38.40	6.78	14.22	41.17	F2:T0;	0%
PR-HN	88.00	<0.01	0.04	0.10	2.30	12.73	39.29	6.46	18.17	45.72	F3: B3	
APQ-h	82.44	<0.01	0.03	0.10	4.21	12.03	38.55	6.46	14.86	42.66		
PR-AI	103.00	0.00	<0.01	0.05	1.97	4.30	24.53	4.47	14.45	75.83		
PR-1C	145.25	0.00	<0.01	<0.01	1.99	3.40	7.51	17.25	35.15	75.19	F1: 90%;	
PR-2C	76.10	<0.01	<0.01	0.09	3.51	8.11	36.10	5.41	10.34	38.43	F2:T0;	0%
PR-HN	81.32	0.00	0.01	0.09	1.92	8.33	36.36	4.44	11.31	39.82	F3: B4	
APQ-h	76.11	<0.01	<0.01	0.09	3.25	8.10	35.99	5.37	10.20	39.02		
PR-AI	113.35	0.00	<0.01	<0.01	1.89	3.09	9.91	4.29	7.29	57.48		
PR-1C	136.58	<0.01	0.00	<0.01	1.91	2.55	5.29	14.77	22.92	57.67	F1: 90%;	
PR-2C	70.90	<0.01	<0.01	0.03	3.58	5.58	23.45	5.44	7.63	25.97	F2:T0;	0%
PR-HN	74.75	0.00	<0.01	0.03	1.86	5.84	23.78	4.23	8.54	27.22	F3: B5	
APQ-h	70.90	<0.01	<0.01	0.03	3.04	5.67	23.52	4.98	7.63	26.51		
PR-AI	170.52	0.00	<0.01	0.05	2.31	5.44	27.84	5.65	15.23	112.20		
PR-1C	239.86	0.00	<0.01	<0.01	2.37	3.81	8.53	26.20	53.05	112.32	F1: 95%;	
PR-2C	113.90	<0.01	<0.01	0.14	4.51	9.93	49.71	6.72	12.34	53.22	F2:T0;	0%
PR-HN	113.72	0.00	0.01	0.14	2.27	10.36	50.22	5.56	14.30	55.79	F3: B0	
APQ-h	113.95	<0.01	<0.01	0.14	4.05	9.88	49.88	6.45	12.27	54.18		
PR-AI	162.56	<0.01	0.02	0.11	2.65	9.78	41.85	7.65	26.42	122.19		
PR-1C	253.61	<0.01	<0.01	<0.01	2.70	4.94	9.79	32.77	72.76	121.96	F1: 95%;	
PR-2C	119.22	<0.01	0.05	0.18	5.52	15.32	61.29	7.83	18.18	63.26	F2:T0;	0%
PR-HN	122.38	<0.01	0.06	0.19	2.59	16.43	62.45	7.55	23.25¡	68.74¡	F3: B3	
APQ-h	119.65	<0.01	0.05	0.18	4.82	15.60	61.64	7.22	20.12	64.85		
PR-AI	164.32	0.00	<0.01	0.10	2.22	5.29	39.62	5.10	19.67	123.38		
PR-1C	244.05	<0.01	<0.01	<0.01	2.27	4.03	10.22	26.21	57.31	122.91	F1: 95%;	
PR-2C	116.62	<0.01	0.01	0.18	4.06	10.65	59.17	6.28	13.16	63.08	F2:T0;	0%
PR-HN	117.40	0.00	0.01	0.18	2.20	10.94	59.72	5.08	14.40¡	64.75¡	F3: B4	
APQ-h	116.87	<0.01	0.01	0.18	3.34	10.77	59.44	5.54	13.91	63.53		
PR-AI	187.94	<0.01	<0.01	<0.01	2.15	3.63	14.07	4.88	8.68	94.22		
PR-1C	226.95	<0.01	0	<0.01	2.17	2.95	6.70	20.76	34.71	94.53	F1: 95%;	
PR-2C	108.41	<0.01	<0.01	0.07	4.17	6.81	37.24	6.25	9.08	40.10	F2:T0;	0%
PR-HN	107.13	0.00	<0.01	0.07	2.06	7.09	37.62	4.81	10.13	41.64	F3: B5	
APQ-h	108.30	<0.01	<0.01	0.07	3.46	6.90	37.37	5.64	9.06	40.76		
PR-AI	160.97	0.00	<0.01	0.06	2.21	5.25	28.90	5.40	15.26	112.07		
PR-1C	230.70	0.00	<0.01	<0.01	2.24	3.63	8.47	27.51	54.68	112.30	F1: 90%;	
PR-2C	107.42	<0.01	0.01	0.15	4.29	9.58	51.13	6.43	11.92	53.84	F2:Tu;	0%
PR-HN	115.53	0.00	0.01	0.16	2.17	10.05	51.71	5.38	13.87	56.25	F3: B0	
APQ-h	107.59	<0.01	0.01	0.15	3.75	9.72	51.30	6.09	12.06	54.69		
PR-AI	158.37	<0.01	0.03	0.13	2.53	9.97	45.46	7.41	27.57	126.60		20%

PR-1C	256.74	<0.01	<0.01	<0.01	2.59	4.79	9.89	36.30	78.57	126.67	F1: 90%; F2:Tu; F3: B3	
PR-2C	244.57	<0.01	0.06	0.21	5.36	15.74	64.96	7.59	18.37	67.47		
PR-HN	251.56	<0.01	0.07	0.21	2.49	16.73	66.37	7.32	23.52	72.77		
APQ-h	126.33	0.02	0.09	0.20	6.77	17.49	57.00	10.71	24.16	62.41		
PR-AI	155.07	0.00	<0.01	0.12	2.12	5.12	41.71	4.84	19.70	122.97	F1: 90%; F2:Tu; F3: B4	0%
PR-1C	235.59	0.00	<0.01	<0.01	2.15	3.89	10.02	27.59	59.45	122.63		
PR-2C	110.37	<0.01	0.01	0.20	3.90	10.40	61.22	5.97	12.79	64.26		
PR-HN	124.77	0.00	0.01	0.20	2.09	10.64	61.57	4.87	13.98	65.85		
APQ-h	110.94	0.00	0.01	0.20	2.93	10.56	61.15	5.76	13.39	64.77		
PR-AI	179.01	0.00	<0.01	<0.01	2.04	3.53	14.73	4.66	8.36	95.07	F1: 90%; F2:Tu; F3: B5	<1%
PR-1C	219.59	0.00	0.00	<0.01	2.07	2.82	6.63	22.23	36.82	95.35		
PR-2C	103.02	<0.01	<0.01	0.08	4.00	6.50	37.97	5.99	8.68	40.73		
PR-HN	108.58	0.00	<0.01	0.08	1.98	6.85	38.26	4.61	9.76	42.19		
APQ-h	102.96	<0.01	<0.01	0.08	3.31	6.65	38.05	5.44	8.68	41.34		
PR-AI	247.60	0.00	<0.01	0.13	2.46	6.44	45.17	6.08	19.69	169.50	F1: 95%; F2:Tu; F3: B0	27%
PR-1C	370.02	0.00	<0.01	<0.01	2.51	4.28	11.19	41.67	86.75	169.89		
PR-2C	1693.08	<0.01	0.02	0.27	4.94	12.10	78.88	7.24	14.77	82.25		
PR-HN	1668.13	0.00	0.02	0.28	2.40	12.59	79.66	6.04	17.01	85.12		
APQ-h	180.28	0.07	0.14	0.20	9.96	22.49	60.53	14.64	28.44	66.53		
PR-AI	522.48	<0.01	0.05	0.22	2.85	13.04	69.98	8.66	37.68	183.51	F1: 95%; F2:Tu; F3: B3	46%
PR-1C	395.08	<0.01	<0.01	<0.01	2.89	5.74	13.51	53.15	118.20	183.64		
PR-2C	2144.24	<0.01	0.10	0.33	6.22	20.84	97.85	8.65	23.82	99.89		
PR-HN	2152.75	<0.01	0.11	0.33	2.77	22.18	99.53	8.49	30.11	106.13		
APQ-h	213.26	0.05	0.15	0.20	8.40	26.42	63.46	14.93	33.36	126.48		
PR-AI	435.44	<0.01	<0.01	0.21	2.37	6.36	66.02	5.47	27.30	184.93	F1: 95%; F2:Tu; F3: B4	48%
PR-1C	378.37	<0.01	<0.01	0.01	2.41	4.66	13.97	41.53	93.64	185.01		
PR-2C	2220.07	<0.01	0.03	0.33	4.48	13.52	95.78	6.76	16.22	99.20		
PR-HN	2196.88	0.00	0.03	0.33	2.32	13.87	96.25	5.45	17.70	100.96		
APQ-h	195.82	0.02	0.14	0.20	7.51	24.39	63.95	13.59	31.18	79.45		
PR-AI	282.31	0.00	<0.01	0.02	2.26	4.02	21.80	5.23	9.98	145.47	F1: 95%; F2:Tu; F3: B5	0%
PR-1C	352.40	0.00	0.00	<0.01	2.30	3.21	8.47	33.03	57.85	145.94		
PR-2C	155.47	<0.01	<0.01	0.17	4.58	7.75	58.18	6.76	10.17	61.02		
PR-HN	153.97	0.00	<0.01	0.17	2.21	8.13	58.57	5.17	11.51	62.75		
APQ-h	155.46	<0.01	<0.01	0.17	3.81	7.81	58.28	6.13	10.13	61.81		
PR-AI	143.36	<0.01	<0.01	0.05	2.17	5.24	26.33	5.27	14.82	97.16	F1: 90%; F2:Tp; F3: B0	0%
PR-1C	204.91	0	<0.01	<0.01	2.20	3.61	8.68	25.79	48.01	97.34		
PR-2C	98.30	<0.01	0.01	0.13	4.16	9.35	45.18	6.23	11.62	47.68		
PR-HN	105.92	0.00	0.01	0.13	2.10	9.79	45.52	5.23	13.55	49.98		
APQ-h	98.42	<0.01	0.01	0.13	3.63	9.43	45.16	5.92	11.70	48.59		
PR-AI	144.13	<0.01	0.03	0.12	2.51	10.10	40.70	7.30	26.31	111.21	F1: 90%; F2:Tp; F3: B3	0%
PR-1C	229.65	<0.01	<0.01	<0.01	2.56	4.85	10.12	33.33	68.43	111.05		
PR-2C	107.58	<0.01	0.06	0.18	5.23	15.26	57.27	7.37	17.90	59.70		
PR-HN	115.07	0.00	0.07	0.19	2.47	16.38	58.33	7.22	22.68	64.77		
APQ-h	108.06	<0.01	0.06	0.18	4.62	15.47	57.62	7.00	18.98	61.28		
PR-AI	139.19	0.00	<0.01	0.10	2.05	5.12	36.84	4.72	18.84	106.74	F1: 90%; F2:Tp; F3: B4	0%
PR-1C	208.15	0.00	<0.01	<0.01	2.10	3.86	10.24	25.17	52.17	106.48		
PR-2C	100.68	<0.01	0.01	0.17	3.79	10.02	53.02	5.81	12.37	56.22		
PR-HN	107.52	0.00	0.01	0.17	2.04	10.27	53.40	4.74	13.52	57.74		
APQ-h	100.84	0.00	0.01	0.17	3.38	10.07	53.09	5.69	12.38	56.82		
PR-AI	157.97	0.00	<0.01	<0.01	2.00	3.44	14.25	4.59	8.27	82.61	F1: 90%; F2:Tp; F3: B5	<1%
PR-1C	195.89	0.00	0.00	<0.01	2.03	2.77	6.76	21.46	34.10	82.85		
PR-2C	94.14	<0.01	<0.01	0.06	3.92	6.41	33.87	5.84	8.58	36.49		
PR-HN	98.99	0.00	<0.01	0.06	1.95	6.66	34.28	4.54	9.60	37.90		
APQ-h	94.05	<0.01	<0.01	0.06	3.29	6.45	34.01	5.33	8.52	37.09		
PR-AI	225.26	0.00	<0.01	0.11	2.43	6.47	40.28	6.01	19.19	151.76	F1: 95%; F2:Tp; F3: B0	29%
PR-1C	329.99	0.00	<0.01	<0.01	2.49	4.31	11.22	37.43	75.17	152.09		
PR-2C	894.61	<0.01	0.02	0.24	4.83	11.81	69.95	7.11	14.42	73.48		
PR-HN	879.47	0.00	0.02	0.24	2.39	12.34	70.55	5.92	16.64	76.28		
APQ-h	159.74	<0.01	0.09	0.19	6.24	18.37	59.93	9.91	24.16	65.86		
PR-AI	212.05	<0.01	0.05	0.19	2.80	12.86	61.11	8.45	34.61	163.66	F1: 95%; F2:Tp; F3: B3	13%
PR-1C	349.20	<0.01	<0.01	<0.01	2.89	5.76	13.36	46.74	101.97	163.77		
PR-2C	1521.67	0.01	0.09	0.29	6.09	19.81	86.10	8.51	22.76	87.91		
PR-HN	1538.46	<0.01	0.10	0.29	2.77	21.06	87.60	8.37	28.66	94.17		
APQ-h	185.39	0.01	0.15	0.20	6.23	24.94	62.74	11.88	30.66	109.67		
PR-AI	216.66	0.00	<0.01	0.18	2.35	6.34	58.32	5.36	26.05	164.97	F1: 95%; F2:Tp; F3: B4	21%
PR-1C	336.36	0.00	<0.01	0.01	2.40	4.65	13.88	37.17	81.93	164.92		
PR-2C	1544.95	<0.01	0.03	0.29	4.41	13.17	84.17	6.61	15.80	87.77		
PR-HN	1552.25	0.00	0.03	0.29	2.29	13.45	84.71	5.33	17.18	89.57		
APQ-h	171.90	0.02	0.11	0.20	7.69	21.04	61.43	13.00	26.68	68.54		
PR-AI	252.86	0.00	<0.01	0.02	2.26	3.97	20.41	5.14	9.84	129.08	F1: 95%; F2:Tp; F3: B5	0%
PR-1C	315.32	0.00	<0.01	<0.01	2.28	3.19	8.62	30.71	51.30	129.42		
PR-2C	142.01	<0.01	<0.01	0.14	4.42	7.53	52.07	6.63	9.94	55.01		
PR-HN	140.50	0.00	<0.01	0.14	2.19	7.91	52.50	5.10	11.29	56.67		
APQ-h	142.01	<0.01	<0.01	0.14	3.71	7.65	52.19	6.03	9.95	55.69		

Table 9.1. Summary of the objective and KPI values of each scenario with the different queue disciplines and the improvement of the optimal APQ-h with respect to the best pure priority rule.

Appendix I Notation and table of acronyms of Part II

SCHEDULING PROBLEM	
Notation	Definition and domain
PARAMETERS	
N	Total number of physicians
P_i	A physician i, $i = 1, \ldots, N$,
M	Number of types of physician groups
G_r	Group of physicians of type r, $r = 1, \ldots, M$,
n_r	Number of physicians of type i, $i = 1, \ldots, M$,
h_r	Workable hours per physician in group G_r over the planning horizon
L	Number of types of shifts
S_j	Group of shifts of type j, $j = 1, \ldots, L$
d_j	Length (hours) of shifts of type S_j
m_j	Number of shifts of S_j in the planning period
γ_{rj}	Denotes whether physicians of type r can work a shift S_j (binary)
T	Number of days for the planning horizon. The planning horizon usually spans a year ($T = 365$)
C	Set of shift characteristics
D_c	Set of types of shifts with characteristics in set C
$\#D$	Number of sets of shifts D_c that generate fairness constraints
δ_c	Minimum number of days between shifts that belong to a set D_c

ν_{1c}	Maximum number of shifts in a set D_c assigned to physicians over a time window of w_{1c} days.
w_{1c}	Time window (days) in which there must be no more than a specific number of shifts from set D_c, ν_{1c}
w_{2c}	Time window (consecutive days) that a physician can work a shift belonging to set D_c
U_c	Average number of shifts in D_c per full-time physician able to work such shifts
W_j	Number of shifts of type S_j that should be worked by each full-time physician eligible to do so
β	Weighting factor in the objective function of the general covering problem

VARIABLES OF THE SCHEDULING PROBLEM FORMULATED AS INTEGER LINEAR PROGRAMMING (ILP) PROBLEM

X_{ijt}	Binary decision variable which determines whether a physician P_i works the shift Sj on day t
H_c^U	Maximum number of hours worked on shifts with characteristics in C by a physician over the planning horizon
H_c^L	Minimum number of hours worked on shifts with characteristics in C by a physician over the planning horizon
J_c^U	Maximum number of shifts in set D_c worked by a physician over the planning horizon
J_c^L	Minimum number of shifts in set D_c worked by a physician over the planning horizon
${J_{rc}}^U$	Maximum number of shifts in set D_c worked by a physician P_i, $i = 1, \dots, N$, of group G_r
${J_{rc}}^L$	Minimum number of shifts in set D_c worked by a physician P_i, $i = 1, \dots, N$, of group G_r

VARIABLES OF THE LINEAR PROGRAMMING MODEL FORMULATED TO SOLVE THE GENERAL COVERING PROBLEM

Z_{rj}	Decision variables: average number of shifts of type $S_j, j = 1, \dots, L$, that should be worked by a physician of type $G_r, r = 1, \dots, M$, in order to cover the demand without exceeding the working hours

F_1, F_2^U, F_2^L, F_3	Deviation variables minimized in the objective function of the covering problem.

GREEDEY RANDOM CONSTRUCTIVE ALGORITHM

$nshifts(t)$	The number of shifts on the t-th day
$LoC(j,t)$	List of Candidates who can be assigned shift S_j on day t
z_{ij}^*	Number of shifts of type S_j assigned so far to physician P_i at the moment of assignment, on day t
Z_{iD_c}	Average number of types of shifts S_j in D_c – the set shift type with characteristics in set C - that should be worked by a physician P_i of type G_r in order to cover the demand without exceeding the working hours $$Z_{D_c} = \sum_{j \in D_c} Z_{ij}$$
$z_{iD_c}^*$	Number of shifts of type S_j in D_c – the set of shift types with characteristics in set C - assigned so far to physician P_i at the moment of assignment, on day t $$z_{iD_c}^* = \sum_{j \in D_c} z_{ij}^*$$
X_i	Set of shifts assigned to physician P_i in the incumbent solution, that is, $X_i = \{\text{shift } j \text{ of day } t \text{ s.t. } X_{ijt} = 1\}$
$g_j(i)$	Greedy function: this is a non-negative definite function. The greater the value of $g_j(i)$ for physician P_i, the greater is his/her need to work this shift j in order to meet the reference values Z_{ij}
$g_{Nj}(i)$	Normalized greedy function
$g_{D_c}(i)$	Greedy function for each of the characteristics affected by the assignment of shift j on day t
$\not{g}_j(i)$	Enhanced greedy function
$p(i)$	The probability of selecting a physician $P_i \in LoC(j,t)$, which depends on his/her value in the greedy function: $$p(i) = \frac{\left(g(i)\right)^\alpha}{\sum_{P_i \in LoC(j,t)} \left(g(i)\right)^\alpha}$$

α	Elitism factor of the greedy algorithm construction phase
VNDS: VARIABLE NEIGHBORHOOD DESCENT SEARCH	
$\rho(X_i, X_i')$	Distance between schedule solutions for a physician
$\max_{depthSearch}$	Maximum depth in the VND
$X' = h_p^k(X)$	Sequence of k shift transfers in which the receiver in one shift transfer is the transferor in the next
$\aleph_k(X)$	A neighborhood of depth k
Q	Maximum number of unfulfilled constraints among all physicians
P_Q	The set of physicians that reach this maximum number of non-fulfillments
$\max_{iter\_VND}$	Maximum number of iterations in the VND
NFO: NETWORK FLOW OPTIMIZATION	
$H_i(X)$	Total working hours
$\bar{H}$	Average working hours
$\max_{iter\_NFO}$	Total iterations of the NFO procedure
LH	Lower limit of the indifference interval
UH	Upper limit of the indifference interval
ε	Factor defining the final window of indifference
$P_{TS}(X)$	Group of transferors
$P_{RS}(X)$	Group of receivers
$P_{IN}(X)$	Group in the indifference interval
WHC	Working hours condition
BSC	Balance shift condition

Appendix J Integer linear programming model: ED physician scheduling problem

Presentation of the full Integer Linear Programming (ILP) model for the ER physician scheduling problem described in Section 6.3, which comprises minimization of the objective function (14) subject to set of constraints (3)-(13).

$$\min \sum_{i=1}^{\#D} (H_{c_i}^U - H_{c_i}^L) + \sum_{i=1}^{\#D} (J_{c_i}^U - J_{c_i}^L) + \sum_{i=1}^{\#D} \sum_{r=1}^{M} (J_{rc_i}^U - J_{rc_i}^L)$$

Subject to

$$\sum_{S_j \in \boldsymbol{S}(t)} X_{ijt} \leq 1 \quad \forall \quad i = 1, \dots, N; \quad \forall \quad t = 1, \dots, T$$

$$\sum_{i=1}^{N} X_{ijt} = 1 \quad \forall \quad S_j \in \boldsymbol{S}(t); \quad \forall \quad t = 1, \dots, T$$

$$\sum_{t=q-\delta_c}^{q} \sum_{j \in D_c} X_{ijt} \leq 1 \quad \forall \quad q = \delta_c + 1, \dots, T; \quad \forall \quad i = 1, \dots N$$

$$\delta_c X_{ij_c q} + \sum_{t=q+1}^{q+\delta_c} \sum_{j} X_{ijt} \leq \delta_c \quad \forall \quad q = 1, \dots, T - \delta_c; \quad \forall \quad i = 1, \dots N$$

$$\sum_{t=q-w_{1c}+1}^{q} \sum_{j \in D_c} X_{ijt} \leq v_{1c} \quad \forall \quad i = 1, \dots, N; \quad \forall \quad q = w_{1c}, \dots, T$$

$$\sum_{t=q-w_{2c}}^{q} \sum_{j \in D_c} X_{ijt} \leq w_{2c} \quad \forall \quad i = 1, \dots, N; \quad \forall \quad q = w_{2c} + 1, \dots, T$$

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} d_j X_{ijt} \leq \rho_r H_c^U \quad \forall \quad i = 1, \dots, N; \quad \forall \quad i \text{ such that } P_i \in G_r$$

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} d_j X_{ijt} \geq \rho_r H_c^L \quad \forall \quad i = 1, \dots, N; \quad \forall \quad i \text{ such that } P_i \in G_r$$

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} X_{ijt} \leq \rho_r J_c^U \quad \forall \quad i = 1, \dots, N; \quad \forall \quad i \text{ such that } P_i \in G_r$$

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} X_{ijt} \geq \rho_r J_c^L \quad \forall \quad i = 1, \dots, N; \quad \forall \quad i \text{ such that } P_i \in G_r$$

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} X_{ijt} \leq J_{rc}^U \quad \forall \quad r = 1, \dots, M; \quad \forall \quad i \text{ such that } P_i \in G_r$$

$$\sum_{t=1,\dots,T} \sum_{j \in D_c} X_{ijt} \geq J_{rc}^L \quad \forall \quad r = 1, \dots, M; \quad \forall \quad i \text{ such that } P_i \in G_r$$

X_{ijt} binary $\forall \quad i = 1, \dots, N; \quad \forall \quad S_j \in \boldsymbol{S}(t); \quad \forall \quad t = 1, \dots, T$

$H_c^U, H_c^L, J_c^U, J_c^L$ integer $\forall \quad D_c$

J_{rc}^U, J_{rc}^L integer $\forall \quad r = 1, \dots, M; \quad \forall \quad D_c$

Appendix K Linear programming model: general covering problem

Presentation of the full Linear Programming model for solving a general covering problem described in Section 7.2 as a basis for the proposed heuristic (Chapter 7). It comprises minimization of the deviation variables introduced in constraints (17)-(24).

$$\min \beta F_1 + (F_2^U - F_2^L) + F_3$$

subject to:

$$\sum_{r=1}^{M} n_r Z_{rj} = m_j \qquad \forall \quad S_j;\ j = 1, \dots, L$$

$$\sum_{j=1}^{L} d_j Z_{rj} \leq h_r \qquad \forall \quad r = 1, \dots, M$$

$$\sum_{S_j \in D_c} Z_{rj} - \rho_r U_c \leq F_1 \qquad \forall \quad r = 1, \dots, M;\ \forall\ D_c$$

$$\rho_r U_c - \sum_{S_j \in D_c} Z_{rj} \leq F_1 \qquad \forall \quad r = 1, \dots, M;\ \forall\ D_c$$

$$Z_{rj} - \rho_r W_j \leq F_j \rho_r W_j \qquad \forall \quad r = 1, \dots, M; \quad \forall \quad S_j \notin \bigcup_c \{D_c\}$$

$$\rho_r W_j - Z_{rj} \leq F_j \rho_r W_j \qquad \forall \quad r = 1, \dots, M; \quad \forall \quad S_j \notin \bigcup_c \{D_c\}$$

$$F_j \leq F_2^U \qquad \forall \quad S_j \notin \bigcup_c \{D_c\}$$

$$F_j \geq F_2^L \qquad \forall \quad S_j \notin \bigcup_c \{D_c\}$$

$$Z_{rj} - \rho_r W_j \leq F_3 \rho_r W_j \qquad \forall \quad r = 1, \dots, M; \quad \forall \quad S_j \in \bigcup_c \{D_c\}$$

$$\rho_r W_j - Z_{rj} \leq F_3 \rho_r W_j \qquad \forall \quad r = 1, \dots, M; \quad \forall \quad S_j \in \bigcup_c \{D_c\}$$

$$Z_{rj} \geq 0 \qquad \forall \quad r = 1, \dots, M; \qquad \forall \quad S_j;\ j = 1, \dots, L$$

$$F_j \qquad \forall \quad S_j \notin \bigcup_c \{D_c\}$$

$$F_1, F_2^U, F_2^L, F_3 \geq 0$$